Spring 5-15-2018

Multi-omics Portraits of Cancer

Kuan-Lin Huang
Washington University in St. Louis

WASHINGTON UNIVERSITY IN ST. LOUIS

Division of Biology and Biomedical Sciences
Molecular Genetics and Genomics

Dissertation Examination Committee:
Li Ding, Chair
Kimberly J. Johnson
Cynthia Ma
Nancy L. Saccone
James B. Skeath
Robert Reid Townsend

Multi-omics Portraits of Cancer

by
Kuan-lin Huang

A dissertation presented to
The Graduate School
of Washington University in
partial fulfillment of the
requirements for the degree
of Doctor of Philosophy

May 2018
St. Louis, Missouri

Table of Contents

List of Figures

Chapter 2: Pathogenic germline variants in 10,389 adult cancers

Chapter 3: Proteogenomic integration reveals therapeutic targets in breast cancer xenografts

Chapter 4: Redefine druggable targets in breast cancer by global phospho-proteomics

Acknowledgments

I am immersed with gratitude.

I thank my PhD mentor Li for providing me ample opportunities to lead exciting projects and mature as a scientist. I also thank my previous mentor Alison for nurturing me to complete my first large-scale genomics project. I thank my thesis committee Jim, Cynthia, Kim, Nancy and Reid for their guidance. I thank our collaborators at the Chen, Carr, Li, Ellis, Fenyö, Payne, Plon, Townsend, Shumlevich and many other labs.

I thank the people who accompany me on this journey. Ding lab members have been remarkable colleagues and friends, including Adam, Reyka, Sohini, Bailey, Chris, Yige, Song, Matt, Jay, Mikes, Wen-weis, Steven, Venkata, Amila, Qingsong, Hua, Dan, Bobo, Clara, Ryan, Shrikar, Jonathan, Lan and Caleb. Thanks to my roommate RC & Cre, DBBS friends including James, Laura, Jeanette, Calvin, Jeremy, Simon, Jerry, Peter and many others.

I thank the patients and their family for making their data available for research. The research in Ding lab is funded by NCI and NHGRI. I also thank the Markey Lucile Pathway and Taiwanese Ministry of Education for scholarships and the DBBS program for supports. Finally, I would like to thank my dad, mom and sister for their unconditional support.

I love you all.

Kuan-lin Huang

Washington University in St. Louis

May 2018

To Those Who Love Life

ABSTRACT OF THE DISSERTATION

Multi-omics Portraits of Cancer

for Arts & Sciences Graduate Students

by

Kuan-lin Huang

Doctor of Philosophy in Biology and Biomedical Sciences

Molecular Genetics and Genomics

Washington University in St. Louis, 2018

Associate Professor Li Ding, Chair

Precision oncology demands accurate portrayal of a disease at all molecular levels. However, current large-scale studies of omics are often isolated by data types. I have been developing computational tools to conduct integrative analyses of omics data, identifying unique molecular etiology in each tumor. Particularly, this dissertation presents the following contributions to the computational omics of cancer: (1) uncovering the predisposition landscape in 33 cancers and how germline genome collaborates with somatic alterations in oncogenesis; (2) pioneering methods to combine genomic and proteomic data to identify treatment opportunities; and (3) revealing selective phosphorylation of kinase-substrate pairs. These findings advance our understanding of tumor biology on a systematic scale and inform clinical practice of cancer diagnosis and treatment design.

Chapter 1: How to paint a multi-omics portrait of cancer

How to paint a multi-omics portrait of cancer? How do we use big data of bio-molecules ranging from DNA, RNA to protein to understand each cancer? And how do we use this knowledge to advance diagnosis and treatment of cancer?

1.1 Genetic predisposition in cancer

A sizable fraction of cancer is heritable[1], yet known common variants explain only a limited percentage of the genetic burden in cancer[2]. More than 100 genes, mostly tumor suppressors, have been found to harbor rare, predisposing alleles[3]. Most reports on germline variants have focused on single cancer types, although mounting evidence has suggested shared predisposition factors across cancer types. Previous pan-cancer studies have highlighted pathogenic germline variants in tumor suppressor genes, including *ATM*, *BRCA1*, *BRCA2*, *BRIP1*, and *PALB2* in adult cancers in The Cancer Genome Atlas (TCGA)[4] and the Collaborative Oncological Gene-environment Study (COGS)[5], as well as *TP53*, *APC*, *BRCA2*, *NF1*, *PMS2,* and *RB1* using 1,120 pediatric cancer cases from the Pediatric Cancer Genome Project (PCGP)[6]. As sequencing projects expand, large-scale, systematic analyses are needed to increase statistical power and to compare predisposition factors among gene categories and cancer types.

Clinical interpretation of germline variants is a pressing challenge. Conflicting claims resulting from variability in sequencing technologies, analysis pipelines, and interpretations hinder the application of such knowledge[7]. Recent American College of Medical Genetics and Genomics–Association for Molecular Pathology (ACMG–AMP) guidelines provide a systematic method for interpretation of sequence variants for genetic disorders[8]; however, a high fraction of variants are relegated to the unknown significance (VUS) category, often due to rarity and conflicting results in existing databases and the primary literature. Systematic analyses of high-throughput data associated with germline variants, such as matching tumor sequencing and mRNA sequencing data, can provide evidence of functional consequence and further inform clinical interpretation. For example, paired normal-tumor sequencing analysis of allele fraction can validate whether variants of tumor suppressors are undergoing positive selection as part of the classic two-hit model[4,9,10] and mRNA analysis can validate whether a germline truncation results in reduced expression. Of note, the current ACMG/AMP guidelines do not make use of this type of somatic analysis evidence for evaluation of germline variants.

1.2 Proteogenomics of cancer xenograft models

Profiling of somatic alteration by next generation DNA sequencing (NGS) has entered clinical practice with the promise of rapid diagnosis of druggable somatic genomic alterations for personalized cancer treatment[11,12]. For example, recent analysis of 4,068 samples from 16 cancer types suggested that repurposing approved drugs based on genomic alterations could provide individualized treatment options for around 40% of tumors[13]. However, clinical evidence for this proposition is limited and has been slow to develop. Further, the signaling and biological effects

of somatic mutations are not routinely determined in human tumor samples even though this is a consideration for rational drug design, response prediction and target prioritization[14,15]. Finally, druggable genomic alterations are not detected in the majority of cases tested by NGS[3]. Comprehensive proteomic analyses provide a potentially valuable approach to validate genomic findings as likely biological drivers and to discover discover opportunities for targeted treatment.

Patient-derived xenografts (PDXs) in immunodeficient mice maintain the histological and molecular heterogeneity of the progenitor human tumor[16] and cytotoxic drug responsiveness is often a transplantable phenotype[17]. Our previous studies have shown that breast tumor PDXs recapitulate major genomic signatures and transcriptome profiles of their original breast tumors[18,19]. Moreover, drug responses to endocrine therapy in breast cancer PDXs resembled that observed in the corresponding patient and endocrine therapy resistance patterns were associated with aberrations in the *ESR1* gene[16]. While comprehensive proteomic characterization of PDX is still lacking, recent studies using reverse phase protein array have identified similar protein profiles between PDX and primary tumors[20,21]. These studies collectively suggest that the PDX approach is a potentially valuable preclinical model for identification and testing of therapeutic targets.

1.3 Proteogenomics of cancer xenograft models

Mutations and alterations in cancer dysregulate kinases and signaling cascades. Large-scale studies of breast cancer have discovered drivers, with genomic and expression changes, in

kinases of the PI3K/Akt signaling and TP53/RB signaling/cell-cycle checkpoint pathways[22,23]. However, genomic findings provide only indirect inference of phosphorylation activity. Furthermore large-scale proteomic studies using Reverse Phase Protein Array (RPPA) are limited with coverage of approximately 200 proteins with available antibodies[24-26]. The impact of candidate driver events on direct signaling are therefore seldom explored in the corresponding tumors. While functional experiments in *in vivo* systems or model organisms enabled controlled assessment of the downstream effects, they need to be complemented by *in vivo* observations that account for the molecular complexity of each tumor.

Mass spectrometry (MS) is evolving rapidly and has cataloged tens of thousands of phosphorylation sites (phosphosites). Recent studies by the Clinical Proteomic Tumor Analysis Consortium (CPTAC) using liquid chromatography (LC) MS/MS have generated proteomic/phosphoproteomic data sets that more deeply profiled the cancer proteome[23,27,28], providing an opportunity to evaluate regulation of phospho-signaling in cancer. Other MS-based studies focusing on the kinome have highlighted kinases that are disrupted in cancer[29], and identified complementary genomic and proteomic alterations in a handful of signaling pathways[30]. However, characterization of kinase-substrate interaction at a single-residue level has been largely limited to *in vitro* and *in silico* predictions[31]. Direct observation of kinase-substrate associations in tumors are required to understand their regulation.

Chapter 2: Pathogenic germline variants in 10,389 adult cancers

2.1 Abstract

We conducted the largest survey of rare germline variants in cancer to date, discovering pathogenic or likely pathogenic variants in 8.9% of 10,389 cases from 33 cancer types. Twenty-one genes showed significant enrichments of these variants, aggregating in heavily predisposed cancers including sarcoma, pheochromocytoma/paraganglioma, and breast/ovarian cancers. The 710 pathogenic or likely pathogenic variants found in tumor suppressors, including truncations of BRCA1, BRCA2, NF1, ATM, and CHEK2, commonly showed low gene expression (49%) and loss of heterozygosity/biallelic events (26%). We also discovered an unexpectedly high number (93) of such variants in oncogenes, including missenses in MET, RET, EGFR and CCND2 duplication frequently associated with high gene expression (39%). Functionalities of variants are further established through co-localization with somatic mutations/post-translational modification sites, family history of carriers and experimental validation of activating RET alleles. Our results provide the largest predisposition landscape in cancer and reveal potential oncogenic mechanisms of germline variants.

2.2 Results

Abbreviation	Cancer	Sample size	Female ratio	Age at onset
ACC	Adrenocortical carcinoma	92	65%	47.2 +/- 16.3
BLCA	Bladder Urothelial Carcinoma	412	26%	68.1 +/- 10.6
BRCA	Breast invasive carcinoma	1076	99%	58.5 +/- 13.2
CESC	Cervical squamous cell carcinoma and endocervical adenocarcinoma	305	100%	48.2 +/- 13.8
CHOL	Cholangiocarcinoma	45	56%	63.6 +/- 12.2
COAD	Colon adenocarcinoma	419	48%	66.7 +/- 13.2
DLBC	Lymphoid Neoplasm Diffuse Large B-cell Lymphoma	41	54%	56.5 +/- 14.3
ESCA	Esophageal carcinoma	184	15%	62.4 +/- 11.9
GBM	Glioblastoma multiforme	393	37%	59.8 +/- 13.6
HNSC	Head and Neck squamous cell carcinoma	526	27%	60.9 +/- 11.9
KICH	Kidney Chromophobe	66	41%	51.5 +/- 14.3
KIRC	Kidney renal clear cell carcinoma	387	36%	60.1 +/- 12.2
KIRP	Kidney renal papillary cell carcinoma	289	27%	61.4 +/- 12.1
LAML	Acute Myeloid Leukemia	142	46%	56.2 +/- 15.4
LGG	Brain Lower Grade Glioma	515	45%	42.9 +/- 13.4
LIHC	Liver hepatocellular carcinoma	375	32%	59.4 +/- 13.5
LUAD	Lung adenocarcinoma	518	54%	65.3 +/- 10
LUSC	Lung squamous cell carcinoma	499	26%	67.3 +/- 8.6
MESO	Mesothelioma	82	18%	63 +/- 9.9
OV	Ovarian serous cystadenocarcinoma	412	100%	59.6 +/- 11.6
PAAD	Pancreatic adenocarcinoma	185	45%	64.9 +/- 11.1
PCPG	Pheochromocytoma and Paraganglioma	179	56%	47.3 +/- 15.1
PRAD	Prostate adenocarcinoma	498	0%	61 +/- 6.8
READ	Rectum adenocarcinoma	145	47%	63.7 +/- 12.2
SARC	Sarcoma	255	54%	60.7 +/- 14.8
SKCM	Skin Cutaneous Melanoma	470	38%	58.2 +/- 15.7
STAD	Stomach adenocarcinoma	443	36%	65.7 +/- 10.8
TGCT	Testicular Germ Cell Tumors	134	0%	32 +/- 9.3
THCA	Thyroid carcinoma	499	73%	47.3 +/- 15.8
THYM	Thymoma	123	48%	58.3 +/- 13
UCEC	Uterine Corpus Endometrial Carcinoma	543	100%	64 +/- 11.2
UCS	Uterine Carcinosarcoma	57	100%	69.7 +/- 9.3
UVM	Uveal Melanoma	80	44%	61.6 +/- 13.9
All	**All 33 cancer types combined**	**10389**	**52%**	**59.2 +/- 14.4**

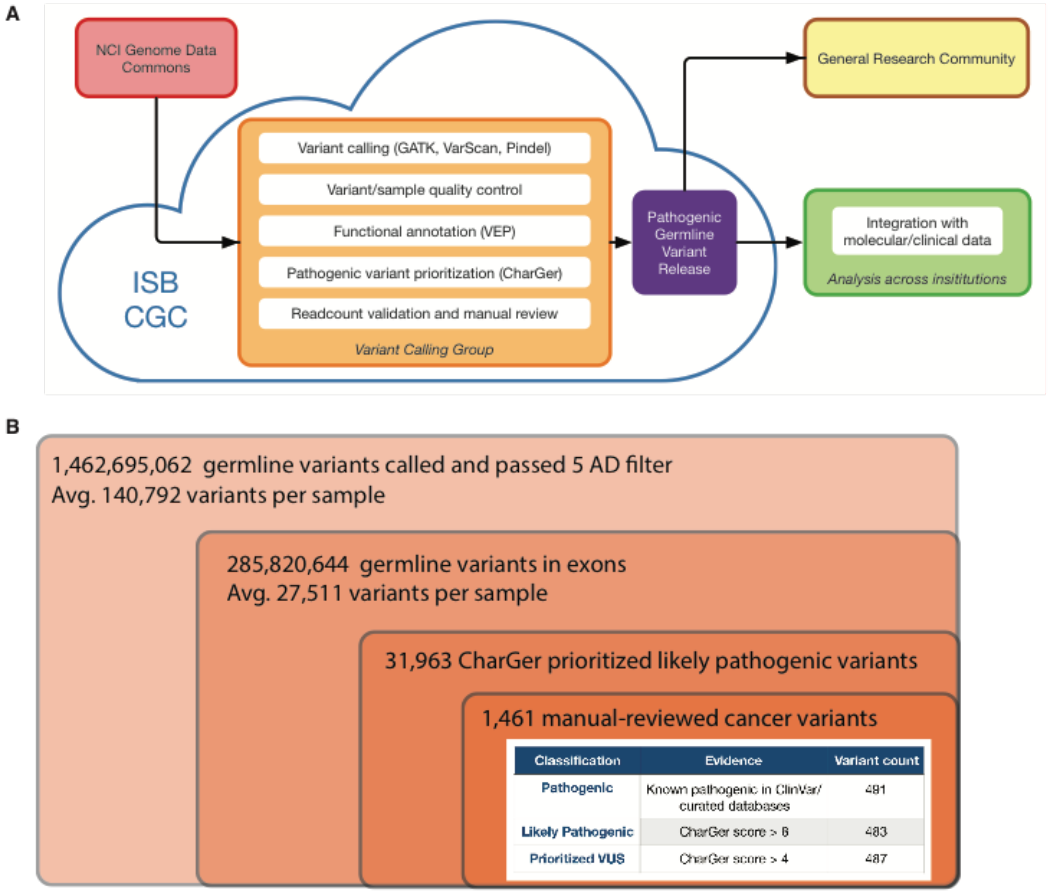


Figure 2.1. Predisposition variant discovery in 10,389 adult cancers of the TCGA PanCanAtlas cohort. (A) A scalable variant calling and data sharing model using ISB Cancer Genome Cloud (ISB-CGC). (B) Number of germline variants at each step of discovery from more than 1.46 billion total germline variants called from WES bam files to 1,461 prioritized, manual-reviewed related to cancer predisposition. 974 pathogenic or likely pathogenic variants are used in downstream analyses.

Data Generation and Sharing on Cloud

The TCGA PanCanAtlas Germline Working Group analyzed germline predisposing variants in 10,389 samples across 33 cancer types. A focus group conducted variant calling on the Institute for Systems Biology Cancer Genomics Cloud (ISB-CGC) and the resulting calls were shared among all investigators for quality control and downstream analyses (**Figure 2.1A**). Specifically, we dockerized the GenomeVIP variant calling system (github.com/ding-lab/GenomeVIP)[32] and deployed more than 121,000 virtual machines running for over 600,000 hours on the ISB-CGC during the course of the project. Variant calls from GATK[33], VarScan2[34], and Pindel[35,36] were merged, filtered, and annotated (**Methods**), resulting in 286,657,499 total exonic variants, ranging from an average of 33,037 exonic variants per individual of African ancestry to 26,640 of European ancestry (**Figure 2.1B**). Our data-sharing paradigm effectively facilitated the analyses required by such an enormous project, avoiding both redundant computation in variant calling/processing and storage of intermediate analysis files in various local computational clusters.

The final set of 10,389 samples passed stringent quality control criteria, showing good coverage, no outlying numbers of variants called, and high concordance with SNP array data (**Methods**). Sample quality control analysis of germline-normal samples revealed an average coverage between 18X and 174X for 151 of 152 predisposition genes known to harbor rare, pathogenic variants (**Methods**). The passed variant calls achieved an average precision above 0.99 when compared to the genotypes obtained through SNP array data. The germline exomes displayed high quality, with an average TiTv value of 2.88 ± 0.17 and lambda value[37] of 0.034 ± 0.003.

The median predicted percent false positive calls across 33 cancer types was less than 5%, ranging from 1.2% (MESO) to 16.1% (KIRC). These resources are shared with the cancer researcher community on the cloud for further evaluation across institutions worldwide.

Pathogenic Variant Discovery across 33 Cancer Types

We developed an automatic variant classification pipeline called CharGer (Characterization of Germline Variants, https://github.com/ding-lab/CharGer) by adopting and extending the ACMG–AMP guidelines[8] specifically for rare variants in cancer. CharGer queries clinical information from ClinVar[38], including variant entry submissions and curated disease-gene associations from OMIM[39], MedGen[40], and Orphanet[41]. We also generated gene-specific databases for known susceptibility genes, including *TP53*, *BRCA1*, *BRCA2*, *RET,* and *TERT* (**Methods**). Further, we extended the published list of 114 known predisposition genes[3], curating a final total of 152 genes that contribute to cancer susceptibility. Overall, each variant is evaluated using data available for any of 12 pathogenic evidence levels and 4 benign evidence tags that contribute to a composite score used for automatic classification. Known pathogenic variants in ClinVar and curated databases are marked as pathogenic, whereas variants with CharGer score > 8 as likely pathogenic, and those with CharGer score > 4 as prioritized VUSs (**Methods**).

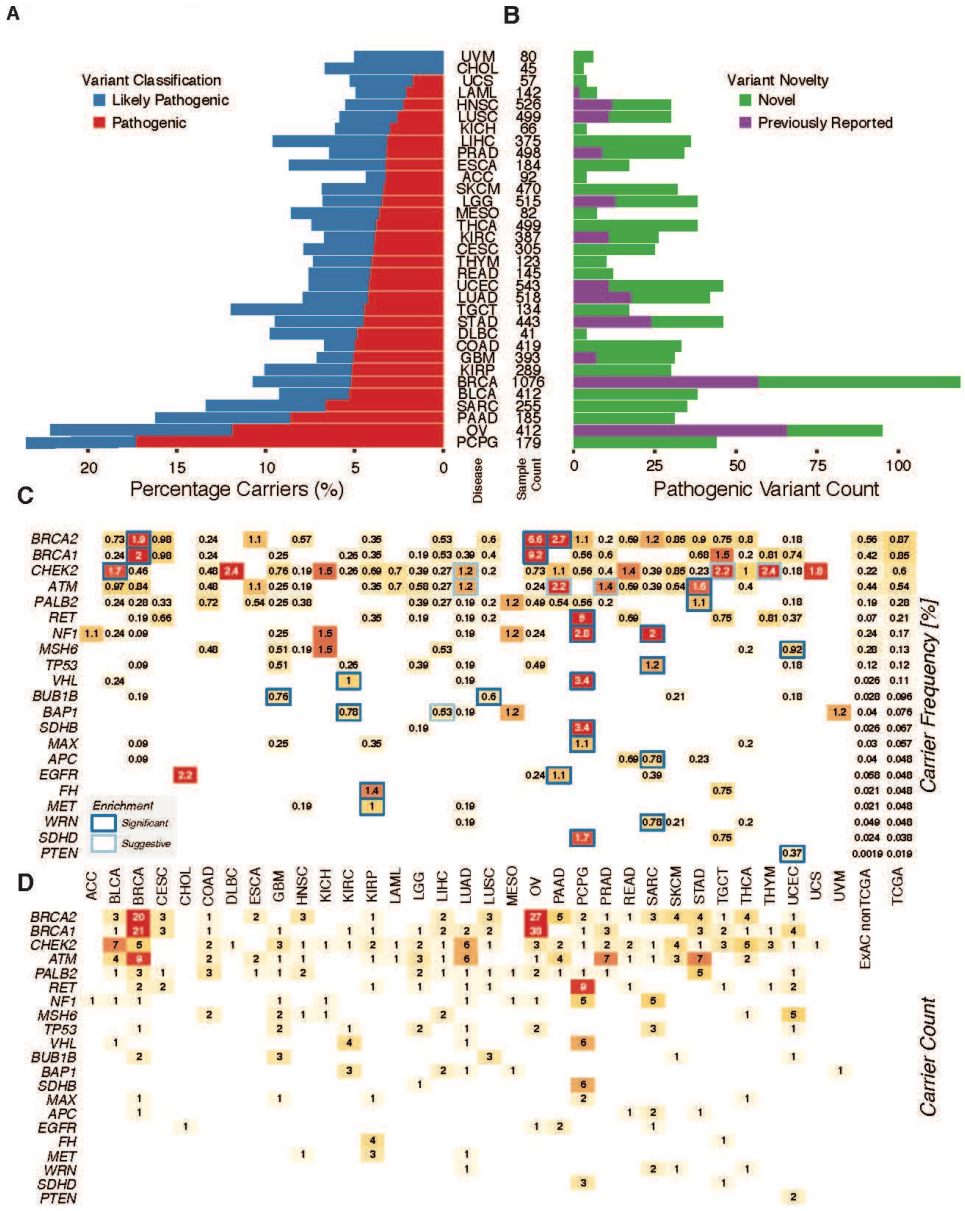


Figure 2.2. Distribution of pathogenic germline variants across genes and cancer types. (A) Percentage of TCGA cases carrying pathogenic and likely pathogenic variants in each of the 33 cancer types. (B) Comparison of pathogenic variants characterized in this study vs. the previous TCGA germline study investigating 12 cancer types (Lu et al., 2015a). (C) Frequency of carriers of pathogenic variants in genes enriched in cancers. Each box indicates the percentage of carriers

of each gene in the specified cancer cohort. The black outlines indicate the cancer type is significantly (FDR < 0.05) enriched for pathogenic variants of that gene. The grey outlines indicate suggestive (FDR < 0.15) enrichment. (D) Counts of pathogenic and likely pathogenic variants in the oncogenes and tumor suppressors enriched in cancers.

We applied CharGer to classify variants found in our selected TCGA cohort into pathogenic, likely pathogenic, and VUS groups. CharGer initially prioritized 31,963 variants in these samples, 1,461 of which were labeled as rare variants (≤ 0.05% AF in 1000 Genomes and complete ExAC r.3.0.1) relevant to cancer, passing manual review in both normal and tumor samples (**Methods**). Combining existing database curation and CharGer results, we classified these into 491 pathogenic variants, 483 likely pathogenic variants and 487 prioritized VUSs (**Figure 2.1C**). This catalog of 974 pathogenic or likely pathogenic germline variants expanded significantly from our previous study, which had focused solely on variants that truncate tumor suppressors in 12 TCGA cancer types[42] (**Figure 2.2B**).

Across all cancer types, 4.6% of cases (n=482) harbored pathogenic variants and another 4.3% (n = 445) carried likely pathogenic variants (**Figure 2.2A**). The frequencies of pathogenic or likely pathogenic variants vary greatly across cancer types, with as expected high rates in OV (22.1%) and BRCA (10.7%). Other cancer types that involve tissue types exposed to environmental factors, such as SKCM (6.8%) and UVM (5%), had lower percentages of carriers. Notably, 23.5% of PCPG, 16.2% of PAAD, and 13.3% of SARC cases carried such variants, suggesting significant contributions of rare germline predisposition in these cohorts[43].

We investigated genes with enriched pathogenic or likely pathogenic variants in each cancer type. Briefly, we first identified cancer types with potential higher enrichment by comparing to pathogenic or likely pathogenic variants identified in the ExAC non-TCGA cohort. We then conducted Total Frequency Testing (TFT)[44] for one cancer type against all other cancer types, subtracting the ones with potential enrichment for each gene (**Methods**). We identified 27 specific cancer-gene associations (FDR < 0.05) and 20 additional suggestive associations (**Figure 2.2C**). Pathogenic or likely pathogenic variants of *BRCA1* and *BRCA2* are highly enriched in OV and BRCA (FDR < 5.51E-05), as expected, while *BRCA2* also showed significant enrichment in PAAD (FDR = 0.022). PCPG is associated with a wide array of predisposition factors, including *RET*, *SDHB*, *VHL*, *NF1*, *SDHD,* and *MAX*. Other genes enriched in multiple cancer types include *BUB1B* in GBM and LUSC as well as *ATM* in STAD, PRAD, PAAD, and LUAD.

On the variant level, we identified 710 pathogenic or likely pathogenic variants in 66 tumor suppressor genes (TSGs) (**Figure 2.2D**). Strikingly, we also discovered 93 pathogenic or likely pathogenic variants in 20 oncogenes, including *RET*, *MET*, *MPL*, *TSHR* and *EGFR*. Twenty-two *RET* variants are found across 11 cancer types. Some variants appear to be cancer specific; for example, all of the 3 pathogenic MET p.H1112R variants are observed in KIRP (papillary renal carcinoma), validating the previously observed co-segregation of the variant in hereditary KIRP[45]. For tumor suppressors, we identified a total of 36 *BRIP1* variants across 22 cancer types. *ATM* and *PALB2* variants are both found in 18 cancer types. In contrast, multiple other tumor suppressor genes showed enrichment in specific cancer types, such as *BRCA1*, *BRCA2* variants

in BRCA and OV (**Figure 2.2C**). For example, all of the 4 tumors containing the pathogenic *BRCA1* p.C61G variants in the ring domain are breast invasive carcinoma.

Two-hit Events

To better understand the biological impacts of the discovered variants, we examined the extent of loss of heterozygosity (LOH) using a statistical test we developed previously[4] (**Methods**). We discovered 170 significant LOH of pathogenic or likely pathogenic germline variants in tumors (**Figure 2.3A**). 22.3% of variants (n=158) in tumor suppressor genes exhibited significantly greater variant allele frequencies (FDR < 5%) in the tumor compared to their corresponding normal sample, indicating partial or full LOH. In contrast, significant LOH is only observed in 6.5% pathogenic or likely pathogenic variants of oncogenes (n = 6), possibly due to their gain-of-function nature and less selection requirement for the activated mutated allele to be homozygous.

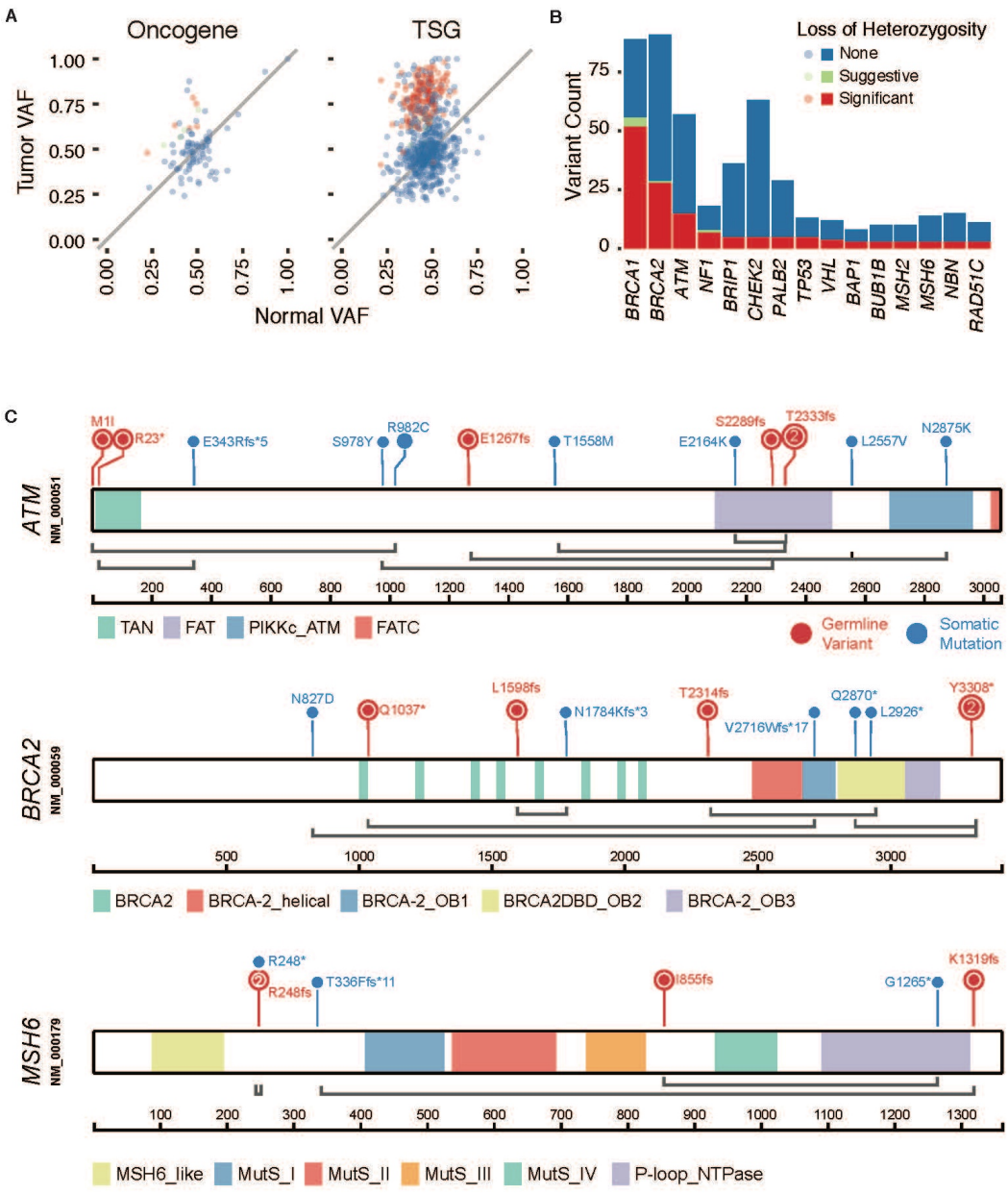


Figure 2.3. Systematic identification of two-hit events in TCGA cancers. (A) Identification of loss of heterozygosity (LOH) in oncogenes and tumor suppressors through comparison of variant allele frequencies in tumor and normal samples. Each dot depicts one variant. The diagonal line denotes neutral selection of the germline variant where the normal and tumor variant allele frequencies (VAFs) are identical. (B) Counts of germline variants showing LOH events in cancer

predisposition genes. Count of variants with significant, suggestive, and no evidence of LOH are shown in red, green and blue, respectively. (C) Candidate biallelic events of pathogenic or likely pathogenic variants coupled with somatic mutations on gene products of ATM, BRCA2, and MSH6. Germline variants are colored in red and somatic mutations are in blue. Coupled germline and somatic events observed in the same case are linked with grey lines.

As expected, strong LOH is observed in cancer types having high hereditary predisposition. The four OV samples containing *BRCA1* p.Q1777fs, p.D825fs, p.W372* and p.E797* each showed highly significant LOH (FDR <= 3.43E-20), whereas *BRCA2* p.E1857fs, p.E294* and p.Y1762* also showed strong LOH in the other 3 OV samples (FDR <= 3.27E-11). The *BRIP1* p.S624* variant showed pathogenic evidence from three independent ClinVar submitters and displayed strong LOH evidence (FDR = 1.31E-16) in an OV sample. Some variants showed evidence of LOH in multiple cancer cases. *MET* p.H1112R, which was previously shown to cause malignant transformation of NIH 3T3 cells[45] in two of the three KIRP samples (FDR = 2.24E-05, 6.98E-3, 0.26, respectively). *RAD51C* p.R193* showed LOH in both BRCA and OV (FDR = 3.04E-12 and 5.79E-05, respectively), but not SKCM (FDR=0.933). The positive selection of these germline variants in the tumor further validates their clinical relevance.

Another manifestation of the two-hit hypothesis is a pathogenic or likely pathogenic germline variant coupled with a somatic mutation in the other copy of the predisposition gene. We identified 45 candidate biallelic events when analyzing the tumor in our cohort (**Figure 2.3B**). Six germline variants of *ATM*, including 2 p.T2332fs and 1 each of p.S2289fs, p.R23*, p.E1267fs and a start loss variant, were coupled with somatic *ATM* mutations in PRAD, READ,

STAD, ESCA, PRAD, and BLCA, respectively. Five cases carrying *BRCA2* germline frameshift truncations, including 2 tumors with p.Y3308*, and 1 each of p.T1598fs, p.A2314fs and p.Q1037*, also harbored *BRCA2* somatic mutations (**Figure 2.3B**). A COAD case carried *MSH6* p.R248fs germline variant/p.R248* somatic mutation that are mutually exclusive in all sequencing reads, clearly supporting the two-hit abruption of both alleles.

Multiple tumor suppressor genes also showed expression patterns consistent with the two-hit hypothesis: an African American KIRP patient with an age onset of 35 carried the pathogenic *FH* p.S187* germline variant and a somatic splice site *FH* mutation and showed low *FH* expression (at 2.07% of KIRP). A BLCA sample carried the *CHEK2* germline p.W93* compounded by 4 different *CHEK2* somatic mutations subsequently showing low *CHEK2* expression (at 1.7% of BLCA). Overall, these results provide supporting evidence of the two-hit hypothesis through LOH and biallelic events of predisposing alleles across tumor types.

Altered Gene Product Expression in Variant Carriers

In addition to expression associated with two-hit events, we systematically investigated the gene and protein expression of each gene in all carriers of pathogenic or likely pathogenic germline variants. Briefly, we calculated the percentile of gene expression for variant carriers relative to other cases in the same cancer cohort. We then conducted a differential expression analysis to look for genes expressed at different levels in variant carriers (**Methods**). We identified 11 significant (FDR < 0.05, Wilcoxon Rank Sum Test) and 12 suggestive (FDR < 0.15) gene-cancer associations (**Figure 2.4A-B**).

In breast cancer, *ATM* (FDR = 5.3E-4, Wilcoxon Rank Sum Test), *CHEK2* (FDR = 3.1E-3), *BRCA2* (FDR = 7.6E-3), and *BRCA1* (FDR = 0.032) carriers all showed significant lower-expression of the respective gene (**Figure 2.4B**). In PCPG, *RET* carriers showed higher RET expression (FDR = 5.3E-4), whereas *NF1* (FDR = 0.013), *SDHB* (FDR = 0.024), *VHL* (FDR = 0.068), and *SDHD* (FDR = 0.07) carriers have lower expression. In addition to BRCA, *ATM* carriers exhibited significantly lower expression in LUAD (FDR = 0.024) and suggestively in PAAD (FDR =0.11) and LGG (FDR = 0.12). We then conducted the same analysis using RPPA data investigating whether the effects extend to the protein/phosphoprotein levels (**Figure 2.4C-D**). Notably, *ATM* carriers were significantly associated with lower protein expression in 5 cancer types, namely, STAD (FDR = 1.2E-3) and PRAD (FDR = 0.031) in addition to validating mRNA expression signals in BRCA (FDR = 3.7E-4), LGG (FDR = 0.031), and PAAD (FDR = 0.048). *CHEK2* carriers also showed lower protein expression of the Chk2 marker in BRCA (FDR = 2.5E-3) and LUAD (FDR = 0.028) and suggestively in TGCT (FDR = 0.15).

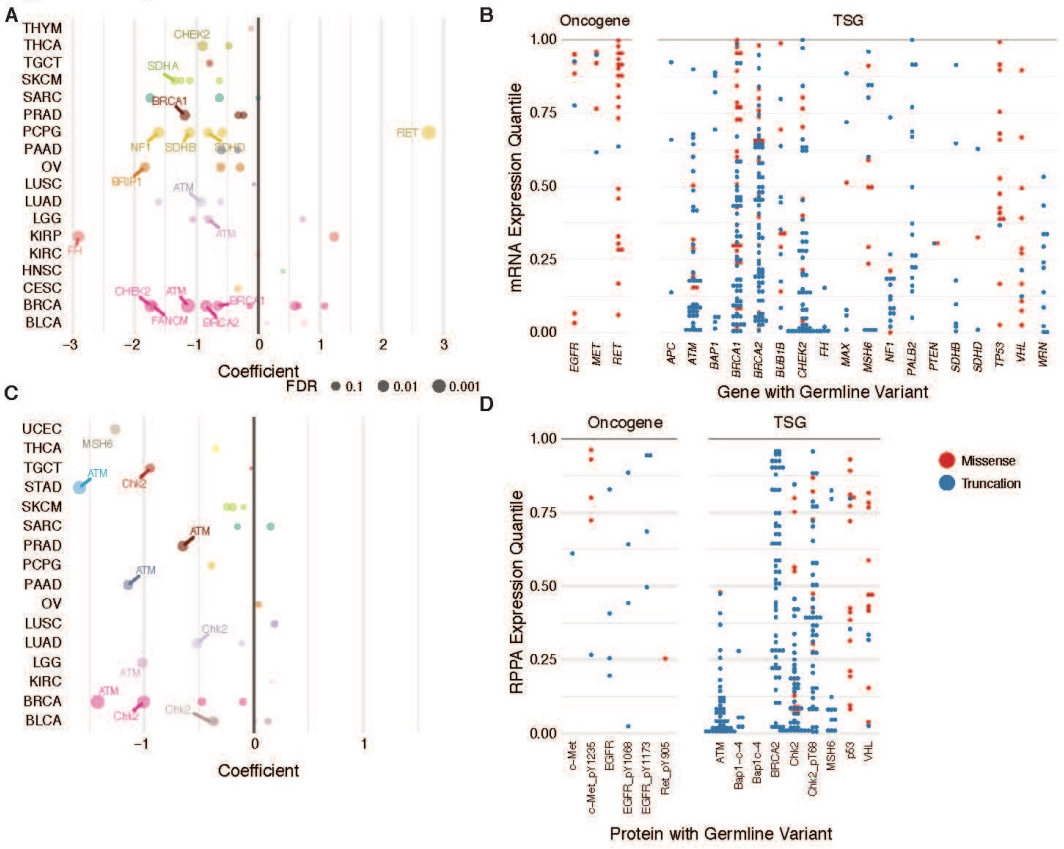


Figure 2.4. Germline variants associated with expression impacts. (A) Volcano plot showing cancer types where the carrier of each gene's germline variant is associated with significantly higher or lower expression of the gene transcript. (B) Distribution of gene expression of pathogenic variant carriers. Each dot corresponds to the gene expression percentile in a case carrying germline variants relative to other cases of their corresponding cancer cohort. (C) Volcano plot showing cancer types where the carrier of each gene's germline variant is associated with significantly higher or lower expression of the RPPA protein/phosphoprotein marker. (D) Distribution of protein/phosphoprotein expression of pathogenic variant carriers. Each dot corresponds to the expression percentile of the RPPA marker in a case carrying germline variants relative to other cases of their corresponding cancer cohort. The genes shown

in (B) and (D) are based on their significant enrichment of pathogenic variants.

Overall, the associated gene expression showed distinct distributions for oncogenes vs. tumor suppressors (**Figure 2.4B**). Pathogenic or likely pathogenic germline variants in tumor suppressors are associated with lower distributions in gene expression than those in oncogenes (Two-sample Kolmogorov-Smirnov test, p = 8.27E-7): 48.5% of such variants in tumor suppressors were associated with the bottom quartile of gene expression compared to 22.9% of variants oncogenes. Further, only 13.0% of the variants in tumor suppressors were associated with the top 25% of gene expression compared to 38.7% of those in oncogenes, suggesting divergent transcriptional regulation of tumor suppressor genes and oncogenes carrying pathogenic or likely pathogenic variants.

On the variant level, tumors with all three *MET* p.H1112R variants were associated with the all showed top 25% *MET* gene expression in KIRP. Notably, 12 cases (including 9 PCPG samples) carrying predisposing *RET* alleles, including p.C618R, p.D631Y, p.C634R/W/Y, p.V804M, p.I852M, p.R912P, p.M918T, showed high *RET* expression in their respective cancer cohorts. The high expression of the variant-associated oncogenes in tumor suggests that cancer cells may preferentially up-regulate pathogenic alleles in these two oncogenes.

Rare Germline Copy Number Alterations

We systematically scanned for rare, germline CNVs in the same 10,389 samples, using both SNP-array data and XHMM analysis as previously described (Fromer and Purcell, 2014; Ruderfer et al., 2016) on whole exome sequence data (**Methods**). We identified 4,050 high-quality CNVs detected using both technologies, affecting ~27% of the cases (**Figure 2.5A**). 20% of the cases have at least one gene impacted by these rare deletions, 10% by duplications. On average, each case had 0.36 deleted genes and 0.83 amplified genes. 44% of the CNVs affected only one single gene while 56% impacted multiple genes (**Figure 2.5B**).

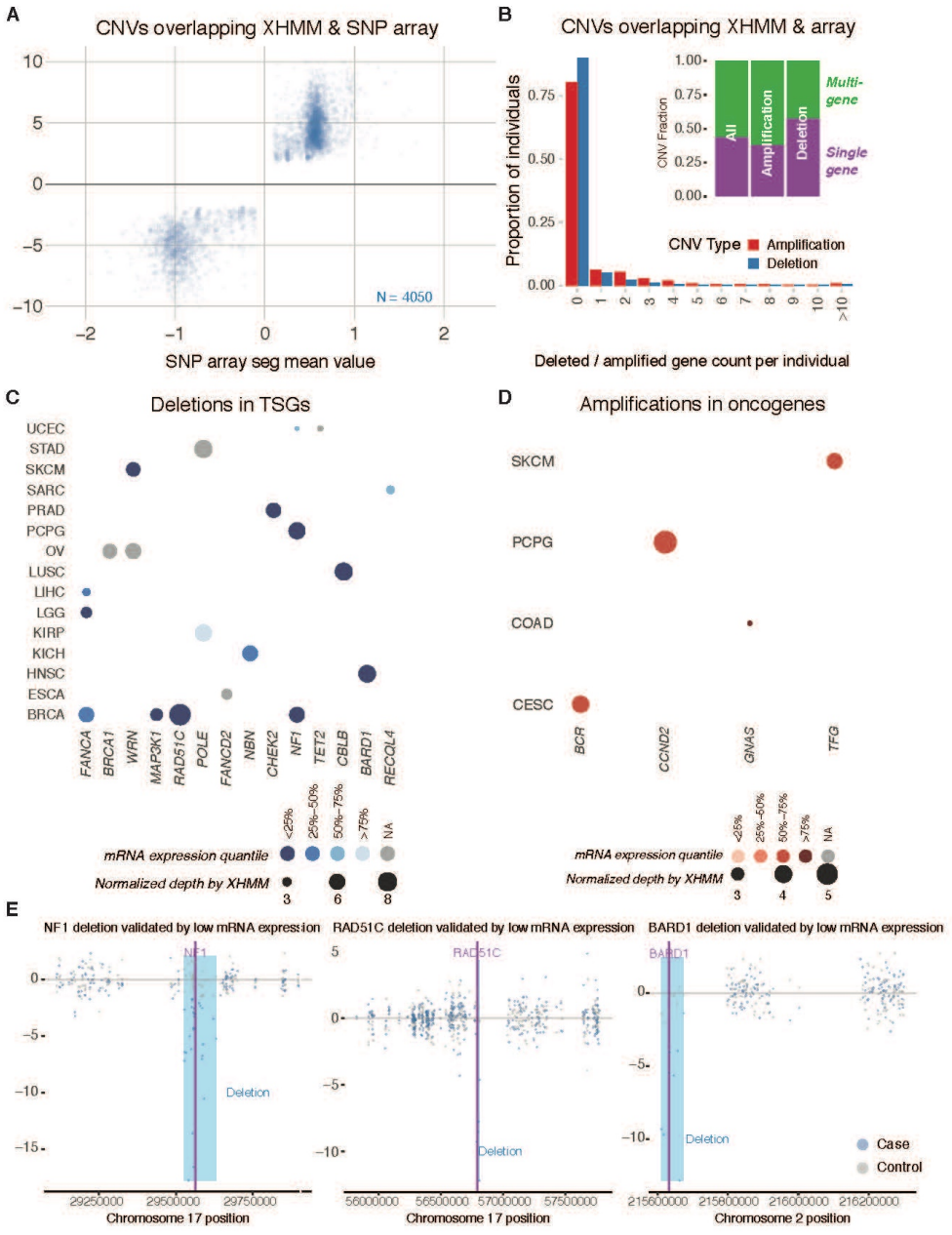


Figure 2.5. Rare germline copy number variations (CNVs). (A) High-confidence Copy Number Variations (CNVs) identified through overlapping events in independent analysis of exome sequencing data (using XHMM) and SNP array data. (B) Characteristics of CNVs identified in the cohort, including fractions of samples carrying deletions/duplications and the number of genes affected by each type of CNVs. (C) Deletions of tumor suppressor genes

associated with reduced expression. (D) Duplication of oncogenes associated with elevated expression. For (C) and (D), each dot represents a CNV event found in a specific cancer type colored according to expression percentile in the cohort. (E) NF1 and RAD51C copy-number deletion associated with reduced expression each in a breast cancer patient case and BARD1 deletion in a HNSC case. The "control" is chosen as another breast cancer or HNSC case characterized in the same whole-exome sequencing dataset without any copy number alterations near this genomic locus.

We then focused on those CNVs that impact our set of 152 cancer predisposition genes and examined their effect on gene expression (**Figure 2.5C-D**). Fifteen deletions of tumor suppressors correspond to the bottom expression quartile in their carriers, while four oncogene duplications were associated with the top expression quartile. Specifically, we found germline deletions of tumor suppressors including *NF1* deleted in two BRCA cases (expression percentiles both at 1.2% and 33%, **Figure 2.5E**) and *RAD51C* deleted in another BRCA case (expression percentile = 1.1%). For oncogenes, an *CCND2* duplication was found in PCGP. A surprising *PIK3CA* duplication was also identified in the blood sample of a PRAD case but not its tumor sample. We subsequently found clinical data indicated a synchronous Non-Hodgkins lymphoma for this case, which is a likely source of *PIK3CA* duplication[46]. Further, a JAK2 amplification was only found in the blood but not normal sample of an OV case representing a possible clonal hematopoietic event. Such examples stress the importance of utilizing both normal and tumor genomics data in discerning germline, somatic and clonal hematopoiesis events.

Independent Genomic Evidence Supporting Pathogenicity

We then sought independent evidence to corroborate the pathogenicity of the identified variants, including (1) significant enrichment in cancer vs. non-cancer cases at a single variant level, (2) co-localization of variants with pathogenic germline alleles found in pediatric cancers or with recurrent somatic mutations and (3) co-localization with post-translational modification (PTM) sites.

To determine whether the pathogenic or likely pathogenic variants are enriched in cancer cases, we conducted association testing by comparing allele frequencies in TCGA cases vs. non-TCGA cases based on the Exome Aggregation Consortium (ExAC r.0.3.1) data of 60,706 individuals [47](**Methods**). We found 175 variants involving 25 genes showing suggestive associations (One-tailed Fisher's Exact test, $P < 0.05$, **Figure 2.6A**), 7 of which passed the multiple testing correction threshold ($FDR < 0.05$) when considering only our previously identified cancer-enriched genes. The top associated variants in tumor suppressor genes include *ATM* p.E1978* (P = 3.50E-06), *BRCA1* p.Q1777fs (P = 2.97E-05), *POT1* p.R363* (P = 3.11E-05) and *PALB2* p.R170fs (P = 5.20E-04). The results also provided supporting evidence of pathogenicity for oncogenic variants such as *MET* p.H1112R (P = 2.00E-03) and *MPL* p.F126fs (P = 0.0161).

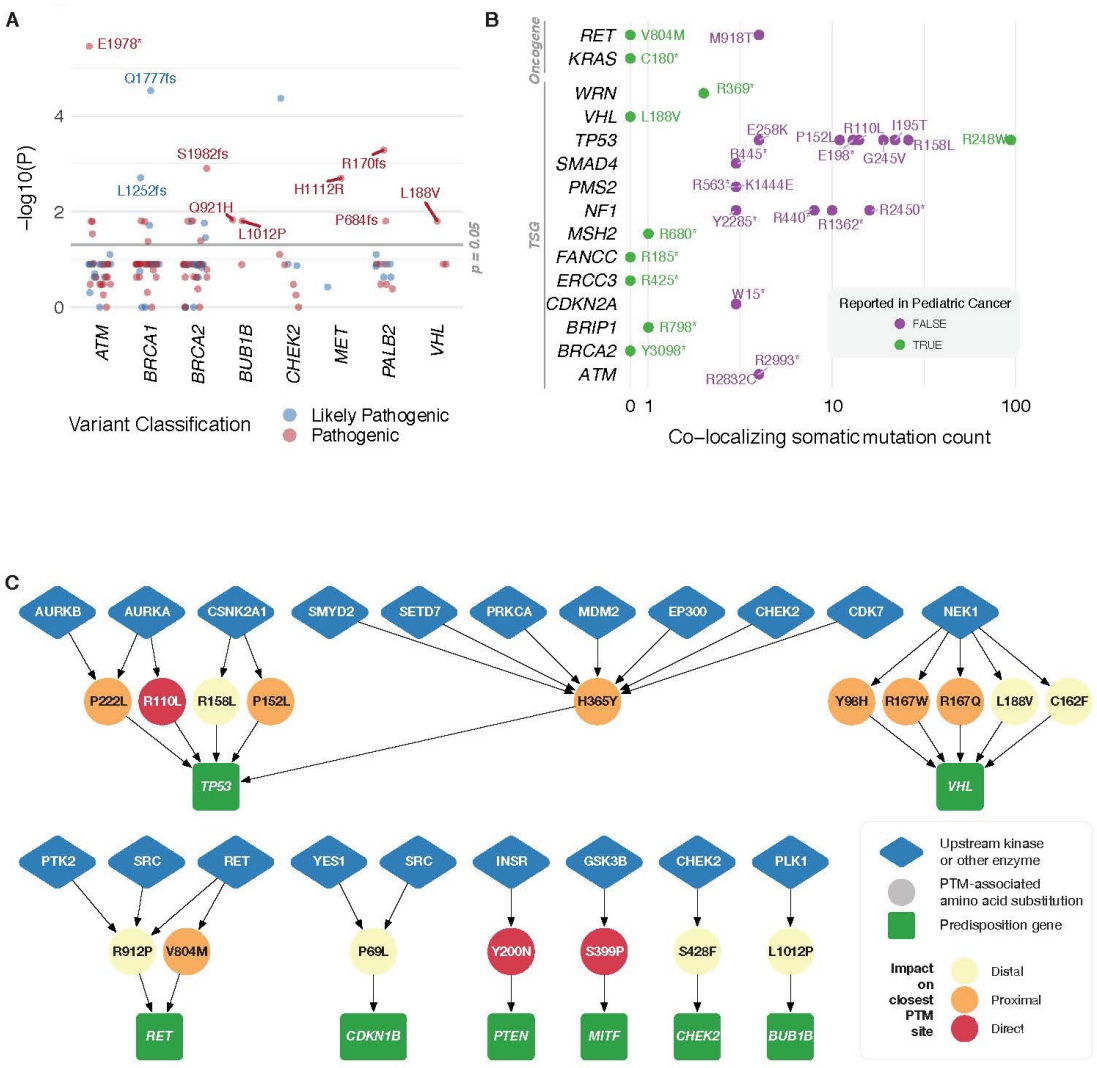


Figure 2.6. Independent evidence supporting functionality of pathogenic variants. (A) Pathogenic germline variants showing significant enrichment in TCGA cases compared to non-TCGA cases in ExAC. (B) Variants with co-localizing recurrent somatic mutations (N ≥ 3 in the TCGA PanCanAtlas MC3 dataset) or pathogenic germline variants in 1,120 pediatric cancers. (C) Site-specific interaction network of predisposition proteins shows how germline substitutions occur in experimentally determined binding sites of upstream kinases and other enzymes.

In our TCGA cohort, we observed 28 pathogenic or likely pathogenic variants previously discovered in 1,120 pediatric cancers[6] (**Figure 2.6B**), including stop-gained variants in *BRIP1*, *ERCC3*, *FANCC*, *MSH2* and *WRN*. Further, we observed 23 incidences of germline variants co-localizing with recurrent ($n \geq 3$) somatic mutations found in the TCGA MC3 cohort (**Figure 2.6B**). Considering unique variants, these include 8 missense variants in *TP53*, 4 *NF1* truncations and 2 *RET* missenses. For example, the *TP53* p.R248W is a highly recurrent somatic mutation (n=94) while being observed as a germline variant in both pediatric rhabdomyosarcoma[6] and LGG. The MEN2-associated allele *RET* p.M918T seen in PCPG and associated with MEN2B disorder was also found as a recurrent somatic mutation (n=4), suggesting shared oncogenic processes in predisposition across pediatric/adult cancers and germline/somatic genomes.

To further evaluate whether this set of 1,461 pathogenic or likely pathogenic variants or prioritized VUSs discovered in TCGA (**Figure 2.1C**) can impact a broader patient population, we examined them in an independent (primarily metastatic) tumor cohort collected at The University of Texas MD Anderson Cancer CenterCancer Center (MDACC), which consists of 3,026 patients in 19 tumor types. Targeted exome sequencing of 200-300 cancer-related genes were previously sequenced from these patients based on an institutional clearinghouse protocol for cancer patients[48]. Thirty unique variants carried by 63 patients were observed in the MDACC cohort from 8 cancer types including breastcolorectal, colorectal, head and neck, and glioblastoma multiforme.

Finally, to investigate additional functional impact of TCGA germline variants on protein signaling, we mapped the pathogenic or likely pathogenic variants to known post-translational modification (PTM) sites in our recently published ActiveDriverDB database[49]. We found that 55 of 132 (42%) protein amino acid substitutions affected PTM sites in 14 genes, significantly more than expected from chance alone (median 8 PTM substitutions expected from 1000 Genomes data; permutation test $P<10^{-5}$).

To illustrate mechanisms of germline variants on signaling networks, we systematically mapped the PTM-associated substitutions to known site-specific enzyme-substrate interactions[49,50] (**Figure 2.6C**). A surprisingly large fraction (17 of 22) of unique substitutions in 8 of 14 genes can be mapped to site-specific interactions with upstream kinases and other classes of enzymes. For example, five substitutions in *TP53* potentially affect binding sites of kinases, such as AURKA, AURKB, and other signaling enzymes, such as MDM2 and EP300. Five *VHL* variants occur in binding sites of the NEK1 kinase that promotes its degradation [51]. *RET* p.V804M and p.R921P potentially affect its auto-phosphorylation sites required for kinase activity [52,53]. Similarly, *CHEK2* p.S428F may affect its auto-phosphorylation [54], suggesting that selected germline variants disrupt and rewire protein signaling networks.

Germline Variants Clustering at Kinase Domains

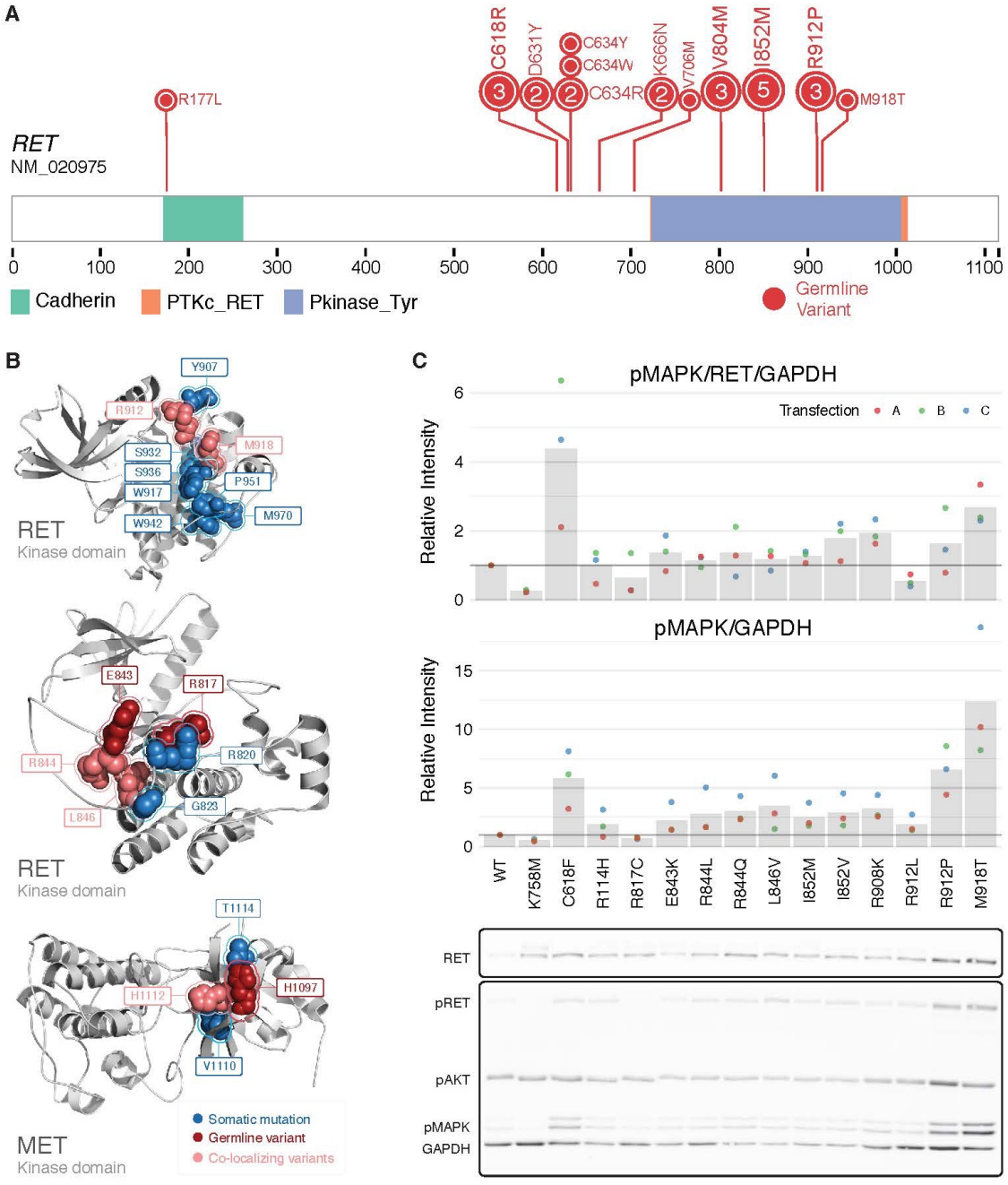


Figure 2.7. Germline variants in the kinase domain of the receptor tyrosine kinase RET. (A) Pathogenic or likely pathogenic germline variants along the RET protein observed in the TCGA cohort. (B) Co-clustering of somatic mutations and germline variants in the kinase

domain of RET and MET shown on 3D protein structures (PDB structures: 21VT, 1R0P, and 1XPD from left to right). Germline variants are colored in red; somatic mutations are colored in blue; amino acid residues affected by both type of mutations are colored in salmon. (C) Experimental assessment of the signaling functionality of RET germline alleles. Ligand-independent RET activity was measured through pMAPK/RET/GAPDH normalized to the ratio observed in wild-type. (D) Experimental assessment of RET germline alleles measured through pMAPK/ GAPDH normalized to the ratio observed in wild-type.

Interestingly, we observed hybrid clusters in the kinase domain of RET: one includes the co-localized germline variants p.R912P/p.M918T and 10 other somatic mutations (**Figure 2.7A-B**) while the other adjacent cluster includes p.I852M along with 5 somatic mutations. Additionally, we also observed germline VUSs co-clustering with somatic mutations in the kinase domain of RET and MET (**Figure 2.7B**) potentially providing additional evidence for pathogenicity. One MET kinase domain cluster centered around residue p.H1112, where the known pathogenic germline variant p.H1112R and the somatic mutation p.H1112Y resides. This cluster contained additional somatic mutations including p.T1114S and the pathogenic p.V1110I and a germline VUS p.H1097R. We further identified a RET kinase domain cluster containing co-localized germline VUSs p.R844L/Q, p.R846V and co-clustered VUSs p.R817C, p.E843K (**Figure 2.7B**), some of which show additional evidence of functionality. For example, RET p.E843K is associated with high expression (97^{th} percentile) and potential enrichment in the cancer population (p=1.7E-4).

Because of the preponderance of variants in *RET* especially in and around the kinase domain, we assessed their functionality by conducting experimental validation of 12 unique germline variants in *RET*, including 3 pathogenic variants and 9 VUSs (Methods). Additionally, we selected a constitutively-active positive control p.C618F (Wells 1994) and a kinase-dead negative control p.K758M.

We assessed the activity of the RET variants by monitoring the downstream pMAPK levels by Western blot in the absence of its ligand GDNF (**Methods**). We first measured RET activity through the ratio of pMAPK/RET/GAPDH (**Figure 2.7C**). As expected, the constitutively-active p.C618F showed ligand-independent activation whereas the kinase-dead p.K758M showed background level of pMAPK. The MEN2 syndrome-associated p.M918T also exhibited higher activity whereas all other germline VUSs found in this study did not show significant change in activity when pMAPK was used as readout.

Activating mutations tend to couple with up-regulation of the oncogenes as seen for RET MEN2 alleles and MET p.H1112R in our cohort (**Figure 2.4B**) and somatic mutations of receptor tyrosine kinases[55]. We thus analyzed the results by measuring RET activity by pMAPK/GAPDH not controlled for the dynamic RET expression (**Figure 2.7D**). While p.R912P was previously shown to co-segregate in familial medullary thyroid carcinoma[56], our results demonstrate that it may also show ligand-independent activation (T-test using pooled SD, unadjusted P = 0.0019). Multiple other variants also showed minor up-regulation of activity and that could be adaptive in a permissive environment and thus warrant further investigations (**Figure 2.7D**).

Association with Ancestry, Onset Age, and Family History

Previous studies have indicated that predisposing variants can show higher frequencies in certain ethnicity groups, seen for example in various rates of prevalence of pathogenic *BRCA1* alterations across Caucasian, Hispanic, African American, and Asian American cohorts[57,58]. We compared rates of pathogenic or likely pathogenic germline variants across ancestries in cancer cohorts when there were at least 25 cases of each ethnicity using the Total Frequency Test (**Methods, Figure 2.8A**). While the power of such analysis is limited at current cohort sizes and no association reached significance after multiple-testing correction, we found suggestive ethnic-enriched predisposing factors: for African cases, *BRCA2* is potentially enriched in LUSC (P = 0.01) whereas *ATM* is potentially enriched in PRAD. In STAD, *BRIP1* and *RECQL* variants are both exclusively found in 2 Asian cases (P=0.054, **Figure 2.8A**).

We then examined whether carriers of pathogenic or likely pathogenic germline variants are associated with a younger age of onset (**Figure 2.8B**). Specifically, we applied a linear regression model to find effects of predisposition genes with more than 3 carriers in each cancer cohort using cancer type as a covariate (Methods). As expected, *BRCA1* variants are associated with younger age of onset in OV (FDR = 7.74E-6) and BRCA (FDR = 0.041) while *BRCA2* variants are suggestively associated in both cancers (FDR < 0.08). We also observed *ATM* variants associating with younger onset age in STAD (FDR = 1.4E-4), *FH* in KIRP (FDR = 0.041) and *VHL* in PCPG (FDR = 0.041).

To identify the predisposition factors in samples with family history, we compared the presence of variants in samples with family history of cancer to those lacking family history of cancer. TGCT and STAD displayed suggestive enrichment of pathogenic or likely pathogenic variants in family history-positive cases (p=0.04 and 0.07, respectively). Interestingly, when including prioritized VUSs, we identified more significant enrichments (p=0.019 and p=0.006, respectively), suggesting predisposition factors in these affected families may exist beyond the strict pathogenic or likely pathogenic variant set.

Out of 522 samples with a positive family history of cancer reported, 50 individuals were found to carry pathogenic or likely pathogenic variants (**Figure 2.8C**). PAAD and TGCT both had 11 samples carrying variants affecting *ATM* (PAAD-specific), *BRCA2* and *CHEK2* among other genes. STAD had 7 samples with rare variants in genes including *CDH1*, a known causal gene for hereditary diffuse gastric cancer, as well as others that are not classically associated with stomach cancer such as *POLE*. Cancers such as COAD, for which hereditary genetics are well described, harbor germline variants in expected genes, such as *MSH6* and *MLH1*.

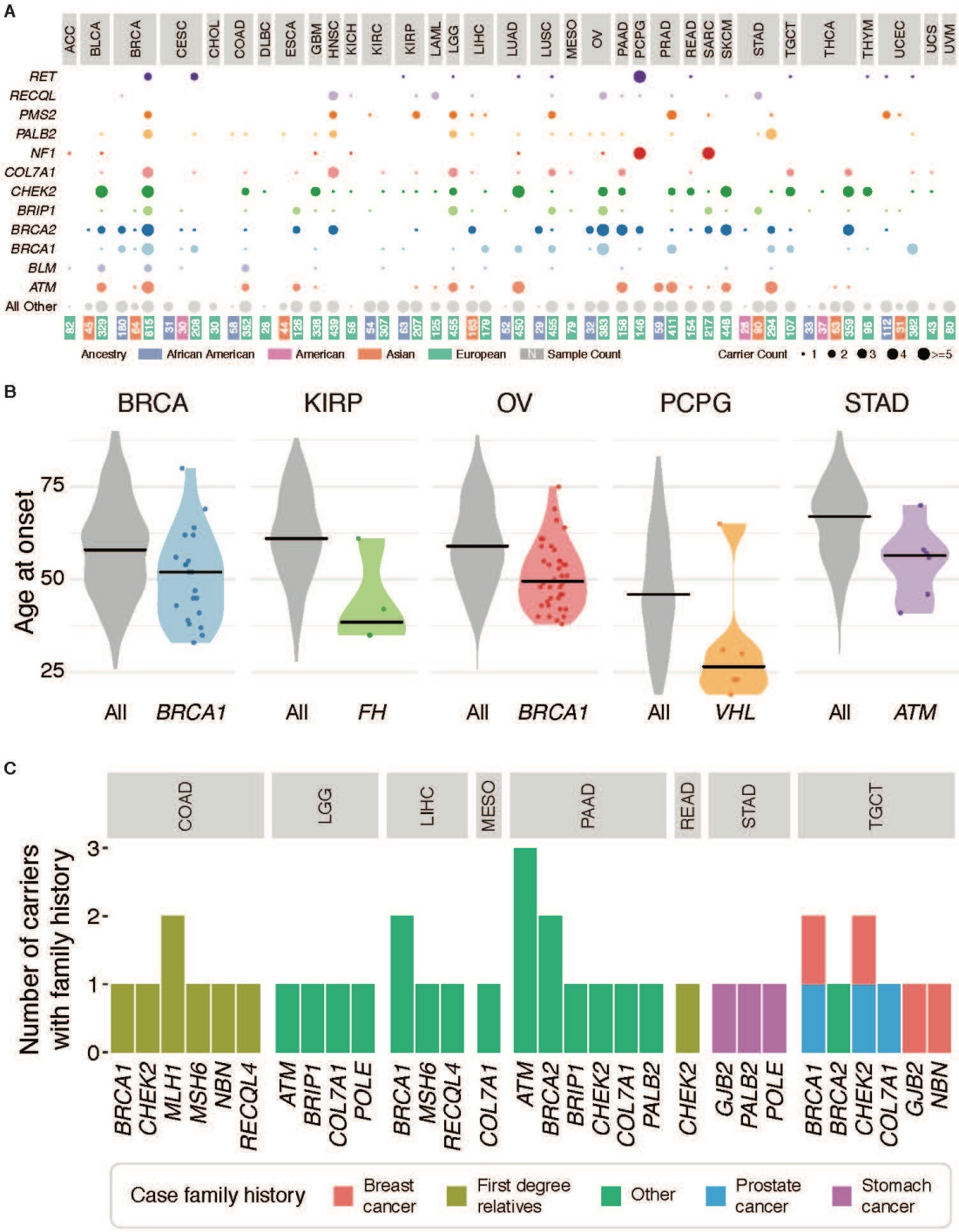


Figure 2.8. Association of ethnicity, onset age, and family history in pathogenic variant carriers. (A) Count of pathogenic variants in cancer cases of each ethnicity. The top bar chart shows the percentage of cases carrying likely pathogenic or pathogenic variants in African

American, Asians, and Caucasians, whereas the bottom bubble plot shows the frequency of specific predisposing factor identified in each cohort with greater than 25 cases. (B) Pathogenic germline variants associated with age at onset in each cancer cohort. The grey violin plot indicates age at onset distributions of non-carriers whereas each of the colored dot represents one carrier of the associated gene in that cancer. (C) Predisposition genes in cases with family history of cancer.

2.3 Discussion

We present the largest catalog of germline variants of cancer to date in 10,389 individuals spanning 33 cancers (**Figure 2.1**). A total of 974 pathogenic or likely pathogenic variants discovered in 8.9% of adult cancer cases, a fraction comparable to recent investigations in smaller cohorts of pediatric and adult cancers[6,59,60]. This comprehensive survey allowed us to establish enrichment of pathogenic or likely pathogenic factors in each cancer (**Figure 2.2**) from *BRCA1/2* in OV and BRCA to *RET/SDHB/VHL/NF1/SDHD* in PCPG. Further, the concurrent systematic discovery of CNVs revealed extremely rare events, such as deletion of *NF1* and *RAD51C* associated with clear expression changes (**Figure 2.5**), suggesting the importance of other germline genomic events beyond SNPs and small insertions/deletions.

Most of the known predisposing factors in cancer are found in tumor suppressors, however an intriguing smaller set of conditions are associated with heritable activating mutations in oncogenes, such as *MET* p.H1112R in hereditary papillary renal carcinoma[45]. In this study, by

conducting the first systematic discovery of germline variants in oncogenes, we identified a striking 93 pathogenic or likely pathogenic variants. Approximately 40% of these variants are associated with high expression supporting their functionality (**Figure 2.4**). Our discovery highlighted oncogenes, including and *TSHR,* as well as genes previously implicated in cancer-related syndromes, but whose contribution to adult cancer remained unclear, such as *ALK* (hereditary neuroblastoma), *RET* (Multiple Endocrine Neoplasia 2A/2B) and *MPL* (Congenital Amegakaryocytic Thrombocytopenia).

Historically, rare pathogenic germline variants have often been overlooked by classification systems due to the lack of evidence in currently available databases and the lack of somatic mutation information in the ACMG/AMP classification system. In this study, we integrated evidence from other omics data to inform variant interpretation practices. In particular, our approach demonstrated the utility of tumor/normal matched sequencing for germline variant interpretation in that they are required to discover two-hit events, including LOH or biallelic events (**Figure 2.3**). Further, by analyzing tumor expression data from RNA-Seq, we identified that approximately half of pathogenic variants were associated with low gene expression in tumor suppressors (**Figure 2.4B**), confirming and expanding findings of germline variants in *BRCA1/2* and *MSH* genes associated with low gene expression[61,62]. The previously-proposed 50 bp rule[63] validated in the gTEX dataset[64] stated that truncations in the last 50bp of the coding sequence are associated with less reduction of gene expression. While our result also confirmed the rule, it does not dictate the lack of expression effect. Thus, careful analysis of gene expression data is required to prove any transcriptional effect imparted by truncations.

We characterized sample and cohort-level interactions between germline variants and somatic mutations in cancer. Within each individual cancer case, we observed that 18% and 4.5% of pathogenic germline variants exhibit significant LOH and biallelic events, respectively (**Figure 2.3C**). At the cohort level, we identified germline variants and somatic mutations affecting the same residues. While these approaches provide systematic evaluation germline variants, careful assessments are required to separate effects from compounding factors, such as passenger somatic copy number events that may induce LOH and non-coding variants affecting transcriptional regulation.

Germline variants overlapping PTMs suggest signaling as a possible predisposition mechanism in addition to PTMs found in other inherited disease mutations[65] and somatic mutations in cancer genomes from TCGA[66]. Beyond the 4 pathogenic variants that directly overlapping PTM sites and additional proximal variants (**Figure 2.6C**), multiple prioritized VUSs also showed potential of modulating PTM. For example, multiple *TP53* variants directly replace arginine residues affected by protein methylation (R110L, R333G, 4x R337H), potentially affecting the target gene specificity of TP53[67]. The *BRCA1* p.Q1281P variant occurs in a known binding site of the ATR kinase that phosphorylates BRCA1 p.S1280 in response to DNA damage[68]. Motif analysis suggests that the substitution replaces a known kinase binding motif of ATR and induces a new motif preferred by cyclin dependent kinases (CDKs) thus potentially resulting in the rewiring of signaling.

Our results also demonstrate the importance of experimental validation of the pathogenic variants. The pathogenic *RET* allele p.R912P showed suggestive ligand-independent activation

(**Figure 2.7**). However, our assay failed to establish functional change in other alleles of familial medullary thyroid cancer including p.I852M (currently with conflicting evidence in ClinVar) and p.R844Q (currently a VUS in ClinVar). It is possible that MEN2-associated alleles exhibit higher expressivity and have easier to detect molecular functional changes with more sensitive assays required to assess weaker gain-of-function alleles of oncogenes that tend to show lower penetrance and their effects may be dependent on cellular context as well. Indeed, additional epigenetics mechanisms such as up-regulation of gene expression may be required for these alleles to achieve their activating potential, such as the candidate germline *RET* VUS p.E843K, which is associated with enrichment in cancer population, high gene expression and conservation among homologs that showed no gain of activity in our assay. Additional modifier genes may also remain to be discovered.

We only identified pathogenic or likely pathogenic alleles in 50 out of 552 cases with family history of cancer and additional predisposing variants may remain to be discovered. For example, none of the samples with family history of testicular cancer displayed a pathogenic variant hit. Further, the STAD and TGCT cohorts showed further enrichment of variant carriers with family history when extending pathogenic variants to include CharGer prioritized VUSs, presenting opportunities for discovery of predisposing alleles beyond the currently-acknowledged pathogenic variants.

Overall, we systematically examined the relationships between functional evidence and clinical classification. The results showed that each germline allele should be carefully evaluated within the relevant context of the somatic genome and downstream expression. The same pathogenic or

likely pathogenic allele often corresponds to similar, yet not identical functional events and phenotypes across individuals. Further, while our results identified potential associations of rare variants with age and ethnicity (**Figure 2.8**), sample sizes of current sequencing cohort limited the power to observe clinical association at a variant level. Integrating omics datasets of even larger cohorts, including those focused on specific cancer types, are thus required to expand the depth of analysis and inform mechanisms of predisposition.

2.4 Methods

Data Generation

Germline variant calling and filtering

TCGA sequence information was obtained from the database of Genotypes and Phenotypes (dbGaP). Sequence data from germline and tumor samples were downloaded by the Institute for Systems Biology Cancer Genomics Cloud (ISB-CGC) from the GDC legacy (GRCh37/hg19) archive. We selected one germline sample and up to one tumor sample per case according to the following procedure. Files designated as TCGA MC3 BAMs were prioritized due to their harmonization. A dockerized version of GenomeVIP was used to coordinate germline variant calling in the guise of integrating multiple tools: Germline SNVs were identified using Varscan (Koboldt et al., 2012) (version 2.3.8 with default parameters, except where –min-var-freq 0.10, --p-value 0.10, --min-coverage 3, --strand-filter 1) operating on a mpileup stream produced by SAMtools (version 1.2 with default parameters, except where -q 1 -Q 13) and GATK (McKenna et al., 2010) (version 3.5, using its haplotype caller in single-sample mode with duplicate and unmapped reads removed and retaining calls with a minimum quality threshold of 10). Germline indels were identified using Varscan (version and parameters as above) and GATK (version and parameters as above) in single-sample mode. We also applied Pindel (version 0.2.5b8 with default parameters, except where -x 4, -I, -B 0, and -M 3 and excluded centromere regions (genome.ucsc.edu)) for indel prediction. For all analyses, we used the GRCh37-lite reference and specified an insertion size of 500 whenever this information was not provided in the BAM header.

All resulting variants were limited to coding regions of full-length transcripts obtained from Ensembl release 70 plus the additional two base pairs flanking each exon that cover splice donor/acceptor sites. Single nucleotide variants (SNVs) were based on the union of raw GATK and VarScan calls. We required that indels were called by at least two out of the three callers (GATK, Varscan, Pindel). In addition, we also included high-confidence, Pindel-unique calls (at least 30x coverage and 20% VAF).

We then further required the variants to have an Allelic Depth (AD) ≥ 5 for the alternative allele. A total of 49,123 variants passed these filters. We then conducted readcount analyses for these variants in both normal and tumor samples. We used bam-readcount (version 0.8.0 commit 1b9c52c, with parameters -q 10, -b 15) to quantify the number of reference and alternative alleles. We required the variants to have at least 5 counts of the alternative allele and an alternative allele frequency of at least 20%, resulting in 31,963 variants. Of these, we filtered for rare variants with ≤0.05% allele frequency in 1000 Genomes and ExAC (release r0.3.1).

We then selected for cancer-relevant pathogenic variants, based on whether they were found in the curated cancer variant database or in the curated cancer predisposition gene list, and their associated ClinVar trait. This resulted in 1,678 variants for manual review using the Integrative Genomics Viewer (IGV) software[69]. For candidate germline variants having the same genomic change as somatic mutations, we further filtered for the germline variants that may have

originated from contaminated adjacent normal samples by eliminating variants called from adjacent normal, the VAF in normal < 30%, and co-localizing with any known somatic mutation. This results in the final 1,461 pass-QC variants for downstream analysis.

We further annotated the corresponding genes of variants as oncogenes or tumor suppressors. We compiled a gene list by combining the oncogenes and tumor suppressors from Vogelstein et al. [70] and the GSEA database (Downloaded 2014-11-25). We removed *NOTCH1* and *NOTCH2* from the oncogene classification in GSEA given their controversial roles. We then further curated several genes, including additional tumor suppressors (*ATR, BARD1, ERCC1, FANCI, FANCL, FANCM, POLD1, POLE, POLH, RAD50, RAD51, RAD51C, RAD51D, RAD54L, MAX*) and additional oncogenes (*AR, STAT3, TERT, MAP2K2*).

Genotype data

We used SNP-array derived genotype data of 522,606 SNPs to infer the ethnicity of each sample. Birdseed genotype files of 11,459 samples were downloaded by ISB-CGC from the Genome Data Commons (GDC) legacy (GRCh37/hg19) archive and converted by us to individual VCF files (github.com/ding-lab/birdseed2vcf) for merging into a single combined VCF file. SNP-array genotypes were also used to assess the precision of germline variant calling in the exome (median precision: 0.99).

Somatic mutation calls

We used TCGA MC3 MAF v3 (updated 17 June 2016) for comprehensive somatic mutation calls across TCGA cancer samples. Specifically, we used mc3.v0.2.8.PUBLIC.maf (www.synapse.org/#!Synapse:syn7214402/files/) (Kyle et al., in review).

Clinical data

We used the clinical data provided by the PanCanAtlas clinical working group (https://www.synapse.org/#!Synapse:syn3241074/files/). For family history information, we used the Clinical data used by the MC3 working group. Ancestry calls of each sample was provided by the PanCanAtlas Ancestry Informative Markers (AIM) working group.

Bioinformatics Analyses

Database curation for variant classification

At the gene level, we extended the 114 predisposition genes compiled by Rahman et al.[3] to a total of 152 genes that contribute to cancer susceptibility based on literature review. At the variant level, in addition to the ClinVar database, we compiled multiple well-curated, gene-specific databases for more comprehensive coverage of known pathogenic variants. These included the IARC *TP53* germline mutation database, NHGRI *BRCA1* and *BRCA2* BIC database (http://research.nhgri.nih.gov/bic/), ARUP MEN2 database for mutations in *RET* (http://www.arup.utah.edu/database/MEN2/MEN2_display.php), and the ASU database (http://telomerase.asu.edu/diseases.html) for *TERT* mutations. We included only the *BRCA1* and *BRCA2* variants marked as clinically important in the BIC database. We also limited our *TP53*

variants to those that were carried by an affected proband and confirmed as a germline variant in the IARC database. We used TransVar[71] and customized scripts to convert all variant entries to standard HGVSg format to ensure proper matching.

Variant classification pipeline and panel review

Briefly, we developed an automatic pipeline termed CharGer (https://github.com/ding-lab/CharGer) to annotate and prioritize variants by adopting the AMP-ACMG guideline. For the automatic pipeline, we defined 12 pathogenic evidence levels and 4 benign evidence levels using a number of datasets, including ExAC and ClinVar (parsed through MacArthur lab ClinVar: https://github.com/macarthur-lab/clinvar), and computational tools including SIFT[72] and PolyPhen[73]. The detailed implementation and score of each evidence level is as follows:

PVS1, PSC1, PM4, PP2, and PPC1: variants in predisposing genes

Variants in the predisposition gene receive one of these evidence level assignments based on variant type and mode of inheritance. Truncations in susceptibility genes that harbor variants with a dominant mode of inheritance are assigned PVS1, but recessive variants in these genes are assigned PSC1. Protein length changes due to inframe insertions or deletions or nonstop variants in genes that harbor variants with a dominant mode of inheritance receive a PM4, whereas recessives receive a PPC1. Finally, missense variants in susceptibility genes are tagged as PP2.

PS1 and PM5: pathogenic peptide changes

Variants that result in identical peptide changes as a previously known pathogenic variant on ClinVar (only those marked as Pathogenic but not Likely Pathogenic) or the compiled list are assigned a PS1. Variants that result in a different amino-acid change at the same position are assigned a PM5.

PM1: hotspot variants

HotSpot3D[74] was run on MC3 somatic mutation calls (hypermutators removed). The protein structure analysis of HotSpot3D identifies mutation clusters, enriched by recurrent and neighboring pockets of mutations. If a germline variant was found to be a somatic mutation with recurrence in at least two samples among all cancer types in a HotSpot3D cluster, then the variant is flagged with a pathogenic characterization of PM1.

PM2 and BA1: minor allele frequency in populations

Variants that are absent or that show extremely low frequency (MAF < 0.0005) in the ExAC dataset are assigned a PM2, whereas common variants (MAF > 0.05) receive a BA1.

PP3 and BP4: in silico analyses

Several ACMG scores use in silico evidence to determine disease association. We used evidence from SIFT[72] and PolyPhen[75], as annotated by VEP[76]. Each in silico analysis was taken as one piece of evidence and if both analyses identified as "damaging" or "deleterious" in SIFT (score <0.05) and "probably damaging" from PolyPhen (score >0.432), the variant was assigned a pathogenic characterization of PP3. Conversely, if both in silico analyses identify in opposition to PP3 characterization (>0.05 for SIFT, <0.432 for PolyPhen), then the variant achieves a

benign characterization of BP4. The score from each fulfilled evidence level is then summed and classified as described in Figure 2.1C.

Burden testing of pathogenic variants

We adapted the Total Frequency Test (TFT)[44] by collapsing pathogenic and likely pathogenic germline variants to the gene level. We then used total allele counts of pathogenic variants identified in the ExAC nonTCGA cohort using the same CharGer classification pipeline for comparison. We deemed one cancer type shows potentially increased burden of a specific gene if the TFT test against ExAC returned FDR < 0.15.

We then tested burden of pathogenic variants for each cancer type and each gene against all other cancer cohorts as controls, subtracting out the cohorts showing suggestive enrichment for the specific gene in the ExAC analyses. Since all our cohorts are called using the same variant calling pipeline, it avoids the potential danger of comparing against ExAC, which was done in a different batch of variant calls. The resulting P values were adjusted to FDR using the standard Benjamini-Hochberg procedure. We subsequently defined significant and suggestive events in terms of FDR thresholds of 0.05 and 0.15, respectively.

Loss of heterozygosity (LOH) and biallelic events analysis

We applied our previously developed statistical analysis method regarding LOH [4] to individually test the missense and truncation germline variant sets. We tested variants in genes carrying

pathogenic or likely pathogenic variants and used variants in other genes to build the null distribution. The resulting P values were adjusted to FDR again using the standard Benjamini-Hochberg procedure. We subsequently defined significant and suggestive events in terms of FDR thresholds of 0.05 and 0.15, respectively.

For biallelic events analysis, we systematically examined the cases carrying both a pathogenic or likely pathogenic germline variant and a missense or truncating somatic mutation in the same gene. The lolliplots are constructed and modified from the PCGP protein paint (https://pecan.stjude.org/proteinpaint) based on the specified RefSeq transcript.

Gene expression analysis

TCGA level-3 normalized RNA expression data were downloaded from Firehose (2016/1/28 analysis archive). The expression percentile of individual genes in each cancer cohort was calculated using the empirical cumulative distribution function (ecdf), as implemented in R. We then used the two-sample Kolmogorov-Smirnov test to compare the expression percentile distribution between variants of oncogenes and tumor suppressors. We also applied the Wilcoxon Rank Sum Test to evaluate the protein/phosphoprotein expression percentile difference between carriers of pathogenic or likely pathogenic variant and non-carriers in cancers where there are at least 3 carriers. The resulting P values were adjusted to FDR again using the standard Benjamini-Hochberg procedure.

To examine the possible location-based effect of truncations, we fitted a linear regression model using expression percentile as the dependent variable and a Boolean indicator to label whether or not the truncation is located at the last 50 base pair of the transcript, controlling for variant classification and truncation variant type.

RPPA analysis

TCGA level-3 normalized RPPA expression data of the tumor samples were downloaded from Firehose (2016/1/28 analysis archive). The expression percentile of individual genes in each cancer cohort was calculated using the empirical cumulative distribution function (ecdf), as implemented in R. We then applied the Wilcoxon Rank Sum Test to evaluate the protein/phosphoprotein expression percentile difference between carriers of pathogenic or likely pathogenic variant and non-carriers in cancers where there are at least 3 carriers. The resulting P values were adjusted to FDR again using the standard Benjamini-Hochberg procedure.

Detection of germline copy number variation events

Whole exome sequencing data on normal samples from 10,389 cases were used for germline CNV detection. XHMM was run as previously described[77]. Base-resolution coverage was calculated by the GATK DepthOfCoverage module (mapping quality > 20) on 209,486 Ensembl coding exon intervals (build GRCh37) retrieved from USCS Table Browser. Exon targets with extreme GC content (> 90% or < 10%) or high fraction of repeat-masked bases (> 25%) or extreme length (< 10bp or > 10kbp) or low mean depth (< 10) were filtered out. The target-by-

sample depth matrix was mean-centered by target dimension. Then principle component analysis was run to remove the systematic bias, where the top 152 components were removed (whose variances were higher than 70% of the mean variances of all components). The resulting depth matrix was normalized to sample-level z-score. During normalization, targets with high variance (standard deviation>50) were filtered out. CNVs discovery was performed using the Viterbi hidden Markov model (HMM) with default XHMM parameters. Quality for each called CNV was calculated by the forward-backward HMM algorithm, as previously described[78].

Array-based CNVs were filtered based on the number of probes (>10), length (>10kb), frequency (<1%), and absolute segment mean value (|log2(copy-number/ 2)| > 0.1). After filtering, the array-based CNV callset consisted of 209,559 CNVs found across 6464 individuals.

Association testing of single variants

We conducted association testing of pathogenic germline variants using a one-tailed Fisher's exact test where the alternative hypothesis assumes the tested variant is enriched in TCGA cases compared to non-TCGA cases in the ExAC data (release r0.3.1). For allele numbers (AN) and allele counts (AC), we used the adjusted counts, where only individuals with genotype quality (GQ) ≥ 20 and depth (DP) ≥ 10 were included. Vcfanno was used to annotate allele frequencies of the germline variants. TCGA allele counts were inferred through subtracting ExAC non-TCGA allele counts from ExAC total allele counts. We conducted the single variant association analysis for all alleles.

Post-Translational-Modification (PTM) site analysis

We mapped the pathogenic or likely pathogenic variants to known post-translational modification (PTM) sites in our recently published ActiveDriverDB database[49], which contains millions of variants and hundreds of thousands of PTMs. We specifically investigated the four most characterized PTM types, phosphorylation, ubiquitination, acetylation, and methylation, all being central to cancer signaling pathways and involved in chromatin regulation and protein degradation switches.

Clinical association analysis (ethnicity and age at onset)

For ethnicity analysis, we again adapted the Total Frequency Test (TFT) by collapsing pathogenic and likely pathogenic germline variants to the gene level. We considered two ethnicities within the cancer type as the cohorts being compared and conducted the analysis between Caucasians and Asian or African Americans or Americans whenever there were more than 25 cases in each cohort.

We used a linear regression model to identify associations between age at onset and germline variant carrier of predisposition genes. We then tested genes with greater than or equal to 3 pathogenic and likely pathogenic variants and 1% carriers in individual cancer cohorts. For both the ethnicity and age at onset association analyses, the resulting P values were again adjusted using the Benjamini-Hochberg procedure.

Family history analysis

Clinical data collected by TCGA projects can provide insight to the family history and/or cancer predisposition of the subject. This data was harmonized across projects using four major fields (Methods). A total of 1,999 patients among 8 cancer types (TGCT, STAD, READ, PAAD, LIHC, LGG, ESCA and COAD) in our cohort have family cancer history information. Across these projects, 522 samples have a family history of cancer. We then used the Fisher's Exact Test to determine whether these cases showed enrichment of pathogenic variant carrier rate compared to cases with family history information yet negative family history.

Co-localizing and co-clustering of somatic mutations and germline variants

We used somatic mutation calls from the TCGA MC3 MAF, defining germline variants located at the same protein residue as recurrent ($n \geq 3$) somatic mutations as co-localizing. We adapted our previously published tool HotSpot3D[74] (v.1.8.0) to conduct co-clustering of TCGA MC3 somatic mutations and pathogenic or likely pathogenic germline variants in genes with available PDB structures.

RET variant function assays

HEK293T cells were authenticated by DNA finger printing targeting short tandem repeat (STR) profiles through Genetica Cell Line Testing. They are negative for mycoplasma as determined by

the absence of extranuclear signals in DAPI staining. Cells were cultured in DMEM (Corning) supplemented with 5% fetal bovine serum (FBS) (Thermo Fisher). Constructions expressing RET variants were generated from a plasmid expressing a wild-type RET (pcDNA3RET9)[79] using Q5 site-directed mutagenesis (New England BioLabs). All constructs were confirmed by sequencing (Supplementary INSERT). Cells were transiently transfected with wild-type or mutant RET constructs using Lipofectamine 2000 (Invitrogen Life Technologies) in six-well plates. Twenty-four hours after transfection, cells were switched to medium containing 0.5% FBS for 24 h before the initiation of 20 minutes of treatment with GDNF (100nM) in a subset of samples. Cells were lysed in buffer containing 20 mM Tris-HCl (pH 7.5), 150 mM NaCl, 1 mM Na2EDTA, 1 mM EGTA, 1% NP-40, 1% sodium deoxycholate, 2.5 mM sodium pyrophosphate, 1 mM β-glycerophosphate, 1 mM sodium orthovanadate, and 1 μg/ml leupeptin (Cell Signaling Technology). Protease and phosphatase inhibitors (Roche) were added immediately before use. Samples (15 ug/lane) were boiled in standard commercial SDS-gel loading buffer and run on SDS 10% polyacrylamide gels. Immunoblotting was performed on Immobilon-P PVDF membrane (Millipore). The following antibodies were used for immunoblotting: rabbit monoclonal anti-phosphor-p44/42 MAPK (Erk1/2) (Thr202/204) antibodies (Cell Signaling #4370S, at 1:1000 dilution), rabbit monoclonal anti-RET (C31B4) antibodies (Cell Signaling #3223S, at 1:1000 dilution), rabbit monoclonal anti-GAPDH antibodies (Cell Signaling #5174, at 1:1000 dilution), rabbit monoclonal anti-phospho-RET (Tyr905) antibodies (Cell Signaling #3221 1:1000 dilution), rabbit monoclonal anti-phospho-AKT (Ser473) antibodies (Cell Signaling #4060 1:1000 dilution), mouse monoclonal anti-RET (C-3) antibodies (Santa Cruz Biotechnologies #sc-365943 1:100 dilution). Appropriate secondary antibodies with infrared dyes (LI-COR) were used, such as donkey anti-rabbit antibodies for the 680nm channel (LI-COR

926-6807) and donkey anti-mouse antibodies for the 800nm channel (LI_COR 926-32212). Protein bands were visualized using the Odyssey Infrared Imaging System (LI-COR) and further quantified by ImageJ.

Code Availability

Analysis codes are available at github.com/ding-lab/PanCanAtlasGermline. GenomeVIP Code for is available at github.com/ding-lab/GenomeVIP. CharGer code is available at github.com/ding-lab/CharGer. Birdseed conversion code is available at github.com/ding-lab/birdseed2vcf.

Chapter 3: Proteogenomic integration reveals therapeutic targets in breast cancer xenografts

3.1 Abstract

Recent advances in mass spectrometry (MS) have enabled extensive analysis of cancer proteomes. Here, we employed quantitative proteomics to profile protein expression across 24 breast cancer patient-derived xenograft (PDX) models. Integrated proteogenomic analysis shows positive correlation between expression measurements from transcriptomic and proteomic analyses; further, gene expression-based intrinsic subtypes are largely re-capitulated using non-stromal protein markers. Proteogenomic analysis also validates a number of predicted genomic targets in multiple receptor tyrosine kinases. However, several protein/phosphoprotein events such as overexpression of AKT proteins and ARAF, BRAF, HSP90AB1 phosphosites are not readily explainable by genomic analysis, suggesting that druggable translational and/or post-translational regulatory events may be uniquely diagnosed by MS. Drug treatment experiments targeting HER2 and components of the PI3K pathway supported proteogenomic response predictions in 7 xenograft models. Our study demonstrates that MS-based proteomics can identify therapeutic targets and highlights the potential of PDX drug response evaluation to annotate MS-based pathway activities.

3.2 Results

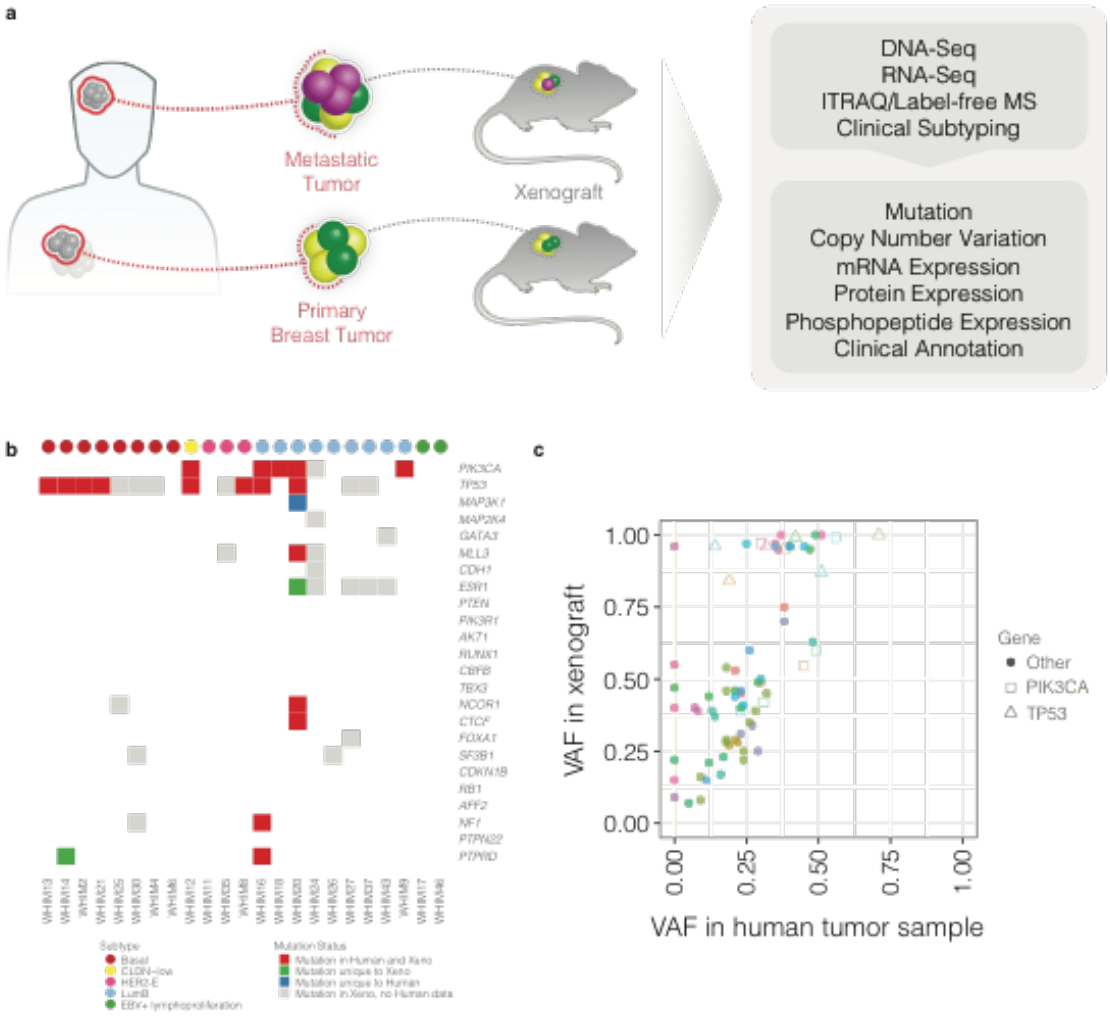


Figure 3.1. Modeling human breast cancer with patient-derived xenografts (n=24). (a) Illustration of generation and proteogenomic characterization of breast cancer xenograft models. (b) Somatic mutations of significantly mutated genes of human breast tumor were recapitulated in xenograft models. Mutation data for 23 WHIMs are shown (exome data were not available for WHIM47). (c) Variant allele fraction analysis showed clonal representation was consistent between human breast tumor and xenografts. Genomic driver events, including missenses and

truncations in *TP53* and *PIK3CA*, were retained in the xenograft models. Each color represents one xenograft sample.

Proteogenomic Coverage of Breast Cancer Xenografts

We selected 24 patient-derived xenograft (PDX) models established from primary or metastatic breast tumors for comprehensive proteogenomic characterization (Figure 3.1a). The human patient cohort was composed of 10 basal, 1 claudin-low (CLDN-low), 9 luminal B, and 4 HER2-Enriched (HER2-E) breast tumors based on PAM50 expression subtyping. We conducted DNA and RNA sequencing respectively for 23 PDX models and in one case Sanger DNA sequencing of hotspot mutations. Isotope Tagging for Relative and Absolute Quantitation (iTRAQ 4-plex)[80] was completed for all 24 PDXs for discovery and Label-Free Quantification (LFQ) for 18 PDXs for validation and confirmation. The comprehensive proteogenomic coverage allowed us to systematically assess the somatic mutation profile, copy number variation (CNV), mRNA expression, protein expression, and phosphosite levels and localizations.

We examined somatic mutations in significantly mutated genes of breast cancer in 12 human tumors and their matching PDXs. Most key somatic mutations in these genes were preserved (Figure 3.1b), validating the genomic fidelity of these PDX models. However, recurrent breast cancer mutations were not detected in two PDX models (WHIM 17 and WHIM 46). While sequencing data were not available for the matched progenitor human tumors, the germline SNPs of human blood normal samples matched with the PDX tumors, validating their patients of origin. Follow-up histological and RNA-seq analyses suggested WHIM17 and WHIM46 are

EBV-positive human lymphoproliferative cells arising in the NOD/SCID/γ mouse strain[81] (Methods).

We then compared the variant allele fractions (VAFs) of exonic somatic mutations in human tumors and derived xenografts when both are available. We found comparable or higher VAFs in xenografts, potentially due to higher tumor purity and selection of some mutant alleles by loss of heterozygosity in PDX models (Figure 3.1c). The relatively high, positive VAF correlation (R = 0.66) implied similarity between human samples and their respective xenografts, consistent with our previous report[19]. Importantly, all *PIK3CA* and *TP53* somatic mutations in the 12 originating human tumors were detected in respective xenografts with comparable or increased VAFs (Figure 3.1c). Among the 22 breast cancer PDX models, 14, including 7 of the 8 basal tumors, harbored *TP53* mutations, 5 luminal B PDXs had *ESR1* mutations, and 5 luminal B and 1 CLDN-low xenografts carried *PIK3CA* mutations.

We quantified mRNA expression of 16,209 unique human genes from RNA-Seq data for these 24 models. We also applied two distinct MS data acquisition approaches (Methods), iTRAQ and LFQ, to quantify the expression ratios of 12,794 human proteins (11,879 genes) and 8,648 proteins (8,035 genes), respectively. A total of 56,874 phosphosites were also confidently identified using iTRAQ. After filtering for observation in at least 10 out of 24 samples (4 out of 9 of the iTRAQ experiments), the relative abundances of 10,069 proteins and 36,609 phosphorylation sites were quantified across tumors by iTRAQ and used in subsequent analyses in this study. The technical replicates in the LFQ (WHIM2 and WHIM16) and the iTRAQ (WHIM13) sets showed high correlation in protein expression levels (R > 0.85). Further,

phosphosite expressions also showed high correlations in technical replicates (WHIM13) of the iTRAQ experiment (R=0.82), validating the technical reproducibility of our proteomic and phosphoproteomic datasets. Of note, while iTRAQ and LFQ quantification were conducted separately and based on different features of the LC-MS data collected from the Q Exactive mass spectrometer (i.e. reporter ion intensity and MS1 peak area, respectively), our analysis showed reasonable correlation between the two measurements after normalization (R = 0.61). The two datasets utilizing the same PDX models but different workflows enabled cross-method validation of global proteomic results.

Integration and Comparison across DNA, RNA, and Protein Data

We first evaluated the correlation between mRNA expression and protein abundance measurements from the iTRAQ experiment (Figure 3.2a) for these 24 PDX models: 83.6% of the genes with sufficient data showed positive correlations with a median Pearson R = 0.536. We investigated whether the trend of mRNA-protein correlations were associated with specific KEGG pathways[82], finding that metabolic pathways involved in house-keeping functions are enriched for genes showing high, positive correlation. For example, genes in the glutathione metabolism pathway showed the highest enrichment for positive correlations (Kolmogorov-Smirnov test, FDR = 5.6e-07). Interestingly, we observed that genes in the ribosome (FDR < 2.2e-16), spliceosome (FDR = 2.0e-13), RNA transport (FDR = 1.7e-5), RNA polymerase (FDR = 1.8e-4), and oxidative phosphorylation (FDR < 2.2e-16) pathways showed relatively lower correlation between mRNA and protein level. These pathways were enriched for genes that do not require translated proteins for their biological functions. Similar pathway-specific pattern of

positive and negative enrichment of correlations were observed by LFQ. The high degree of mRNA-protein correlation observed in the PDX samples is consistent with recent results obtained for human breast tumors[83] and colorectal cancer[84], suggesting that PDXs closely mimic the respective human breast tumors in their relationship between mRNA and protein.

We then examined the correlation of CNV, mRNA, and protein expression levels for several key genes for breast cancer biology: *EGFR*, *ERBB2* (*HER2*), *ESR1*, *GATA3*, *PGR*, *PIK3CA*, *AKT1/2/3*, *MTOR*, and *TP53*. In most cases, we observed consistent relationships between CNV, mRNA, and protein expression levels. Compared to PDXs of non-luminal subtypes, luminal B breast cancer xenografts, as expected, showed higher mRNA and protein expression of *ESR1* and *PGR*, consistent with their positive ER and PR status (Figure 3.2b). 5 out of the 6 *PIK3CA* mutations were observed in luminal B PDXs, which tended to also show higher mRNA and protein expression levels of *GATA3*. In contrast, a larger proportion of basal PDXs expressed higher protein levels of EGFR. Strong HER2 expression at the mRNA, protein and phosphoprotein levels were detected in WHIM8 and WHIM35, both derived from HER2-positive breast cancers. Overall expression patterns for key genes were consistent with the clinical subtype diagnosis across CNV, mRNA, protein, and phosphosite analyses.

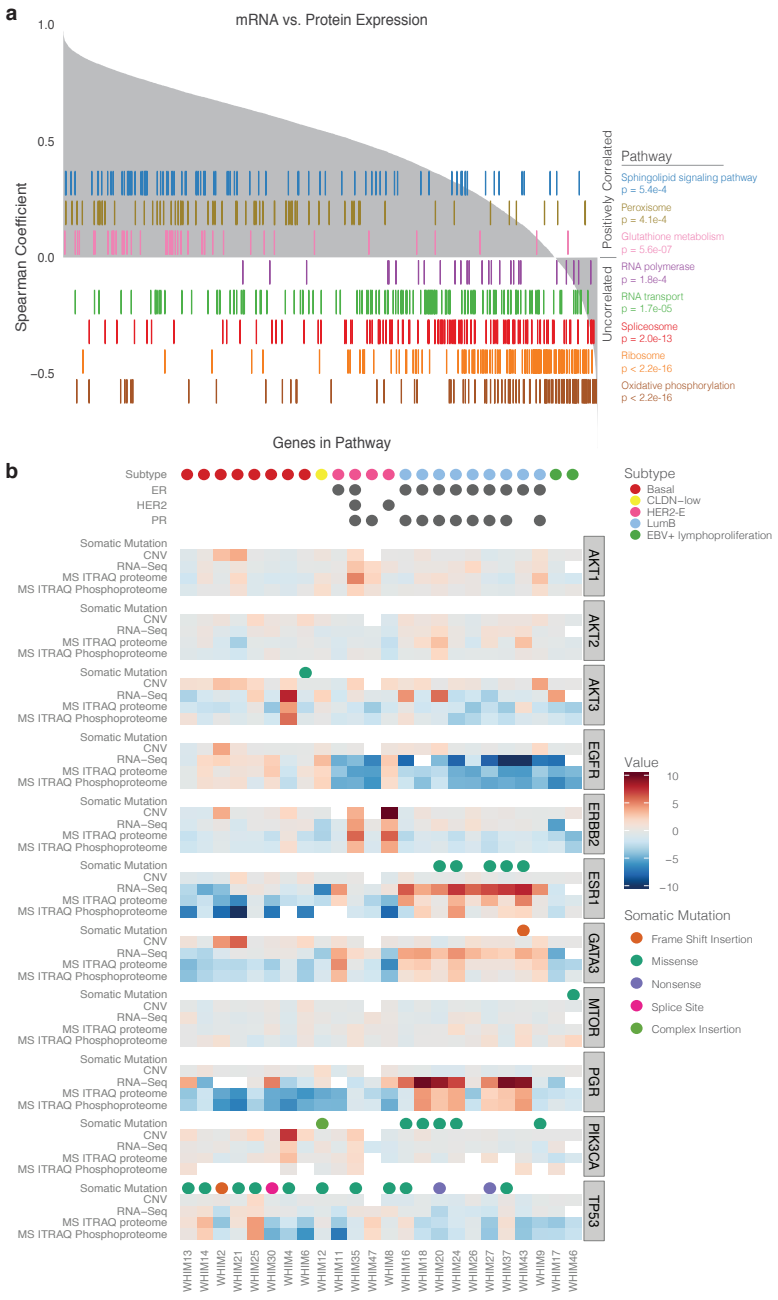


Figure 3.2. Correlation analysis across DNA, RNA and protein levels in PDX samples. (a) Correlation between mRNA and iTRAQ protein expression levels identified pathways with significantly concordant or discordant mRNA-protein expressions. Genes were aligned along the x-axis by the rank of their Spearman correlation coefficient between mRNA and protein

expression levels. Each color represents one significantly associated pathway, and each bar represents one gene in the pathway. (b) Proteogenomic summary of xenograft shows relationships among mutation, CNV (normalized log-R ratio), mRNA (log-transformed and normalized RSEM values), proteomic (normalized log2 ratio to reference), and phosphoproteomic expression (normalized log2 ratio to reference) levels of breast cancer-related genes in 24 PDX samples across 4 intrinsic subtypes. Expression values from each dataset were calculated as described (Methods) and truncated to a maximum of 10 and a minimum of -10 for visualization.

Proteomic Subtyping of Xenografts and Human Breast Tumors

Molecular subtyping of breast cancer based on mRNA expression profiles has been shown to correlate with prognosis and has treatment implications[85,86]. As expected, transcriptome clustering based on PAM50 genes from RNA-Seq data largely reproduced intrinsic subtypes of the breast tumors (Figure 3.3a). To explore proteomic and phosphoproteomic subtype classifications in xenograft models, we conducted unsupervised clustering of proteomes and phosphoproteomes of the 24 xenograft samples based upon the top 436 variably expressed proteins showing a standard deviation greater than 2 from the iTRAQ proteome. Two distinct clusters emerged: one contained all basal tumors and the only CLDN-low tumor (WHIM12), while the other included all luminal B and HER2-E breast tumors; HER2-E tumors did not show a proteomic expression profile distinct from luminal B samples (Figure 3.3b). Clustering analysis using the same gene markers from the transcriptome and the LFQ proteome further supported the separation into these two proteomic subtypes, although the minor differences between the transcriptome and the proteome clustering suggested distinctions between mRNA and protein

levels. The proteomic subtypes defined by the top 968 most variably expressed mouse host proteins did not segregate based on luminal and basal subtypes. Besides differing in mRNA expression, luminal and basal breast cancer PDXs also showed consistently distinct proteomic expression profiles, supporting their distinct biological origins.

We then utilized the iTRAQ phosphopeptide expression data to infer phosphoproteomic subtypes. 1,737 unique phosphosites with standard deviation greater than 2.5 were included to conduct hierarchical clustering (Figure 3.3c). These analyses of the phosphoproteomic data produced two major clusters segregating the luminal B and basal subtypes. Again, the WHIM37/WHIM47/WHIM26 group and the WHIM17/WHIM46 group, as observed in the proteome clustering, grouped closely together. Gene and protein expression of lymphoid lineage markers showed high expression of CD20 and JAK3 in WHIM17 and WHIM46 (Methods), consistent with their positive EBV status and histological diagnosis as human lymphoproliferative cells arising in an immunocompromised mouse background. Overall, while the basal and luminal clusters remain consistent, the hierarchical distances between PDX samples within the two major clusters differ between data types. The departure of proteomic and phosphoproteome subtypes from mRNA expression-defined subtypes suggests independent layers of molecular heterogeneity provided by distinct proteomic analyses.

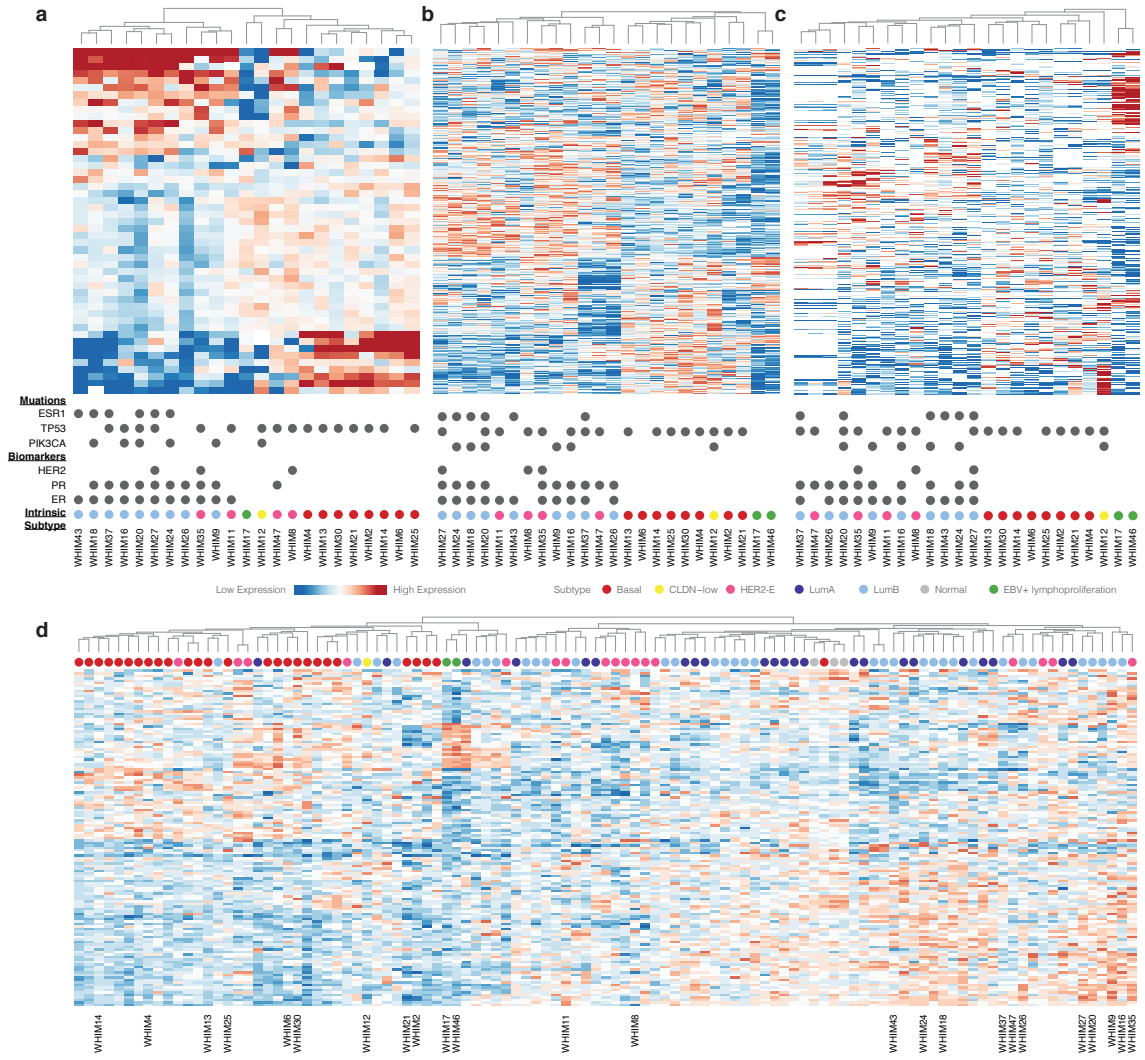


Figure 3.3. Unsupervised hierarchical clustering of breast cancer PDX transcriptomes, proteomes, phosphoproteomes and combined PDX and human breast tumor proteomes. (a) Transcriptomic clustering of PDX breast tumors based on the PAM50 gene expression markers. (b) Proteomic clustering of PDX breast tumors based on the top 436 variably expressed proteins. (c) Phosphoproteomic clustering of PDX breast tumors based on the top 1,737 variably expressed phosphosites. (d) Proteomic clustering using only 133 non-differential expressed proteins between WHIM and human breast tumor samples. The clustering reproduced the basal-

enriched and luminal-enriched clusters, where PDX (n = 24) and TCGA human breast tumor samples (n = 77) cluster based on their subtypes. The non-differentially expressed proteins were identified through a t-test with FDR > 0.3 between the PDX and the TCGA human tumor samples. The PDX tumors are labeled by their WHIM IDs whereas the human tumors are not labeled.

We then investigated whether human proteomic subtypes could be recapitulated in PDX models by including an additional 77 proteomes from TCGA human tumors that were processed concurrently in the iTRAQ experiment[83] (Methods). To reduce clustering bias imbued by the mouse contribution to the proteome in the PDX samples, we excluded proteins showing differential expression between human tumor and xenograft (t-test, FDR ≤ 0.3). Additional requirements, including the presence of detectable protein in more than 10 samples and minimum difference of 2 standard deviations in the merged proteome, resulted in 133 proteins qualifying for un-supervised clustering. Consistent with the subtyping analysis of the 24 PDXs alone, we identified two major clusters: one that included all but one basal breast tumor and the other comprised mostly of luminal tumors (Figure 3.3d). Similar to the proteomic subtypes of the PDX cohorts, luminal tumors and HER-2 tumors did not show clear separation, although several sub-clusters were identified. Importantly, xenograft proteomes clustered adjacently to the human proteomes of their respective subtypes, validating the fidelity of basal and luminal proteomic signatures discovered in PDXs.

To search for defining markers between the basal and luminal B subtypes, we conducted differential expression analysis between the PDXs of the respective subtypes. We found several proteomic markers that were differentially expressed in both the LFQ and iTRAQ datasets (t-test, FDR ≤ 0.05), including SPR, GSTP1 and SERPINB5. We then conducted gene-set enrichment analysis based on the Reactome pathway database[87] to investigate patterns of differential expression (Methods). The basal subtype breast tumors were up-regulated in most of the significantly differentially-expressed pathways (FDR ≤ 0.01) identified by both LFQ and iTRAQ datasets, including extracellular matrix organization, cell cycle, and collagen formation. In comparison, the luminal B breast tumors showed higher expression in genes related to organelle biogenesis and membrane trafficking.

Activated Pathways Revealed by Phosphorylation Profiles

Cancer driving somatic events trigger major changes in downstream signaling to launch the tumorigenic cascade[88]. To search for tumor-specific activated pathways in PDX tumors, we systematically evaluated and compared phosphoproteome profiles of gene sets from KEGG signaling pathways (Methods). Phosphorylation enrichment analysis identified 12 significantly activated pathways, including Ras, MAPK, and NFκB signaling, in 4 xenografts (FDR ≤ 0.01). WHIM9 exhibited elevated phosphorylation of the MAPK signaling pathway (FDR = 9.69e-6, Figure 3.4a). Interestingly, WHIM9 carried a recurrent somatic mutation, *KRAS* p.A146V[89,90], which may have driven canonical MAPK pathway activation. Further, WHIM12 exhibited activation of the Ras signaling pathway (FDR=4.28e-5), along with an outlier protein expression

of MET, a receptor tyrosine kinase upstream of the Ras signaling pathway (Figure 3.4b). Interestingly, WHIM17 and WHIM46, both harbored BTK and PLCG2 protein overexpression, exhibited overall high phosphorylation of the NFκB signaling pathway (FDR = 6.94e-3, 4.53e-5 respectively). This observation further supports the strong similarity between these two PDX models based on protein/phosphoprotein clustering (Figure 3.3b,c) and their classification as EBV-positive lymphoproliferation.

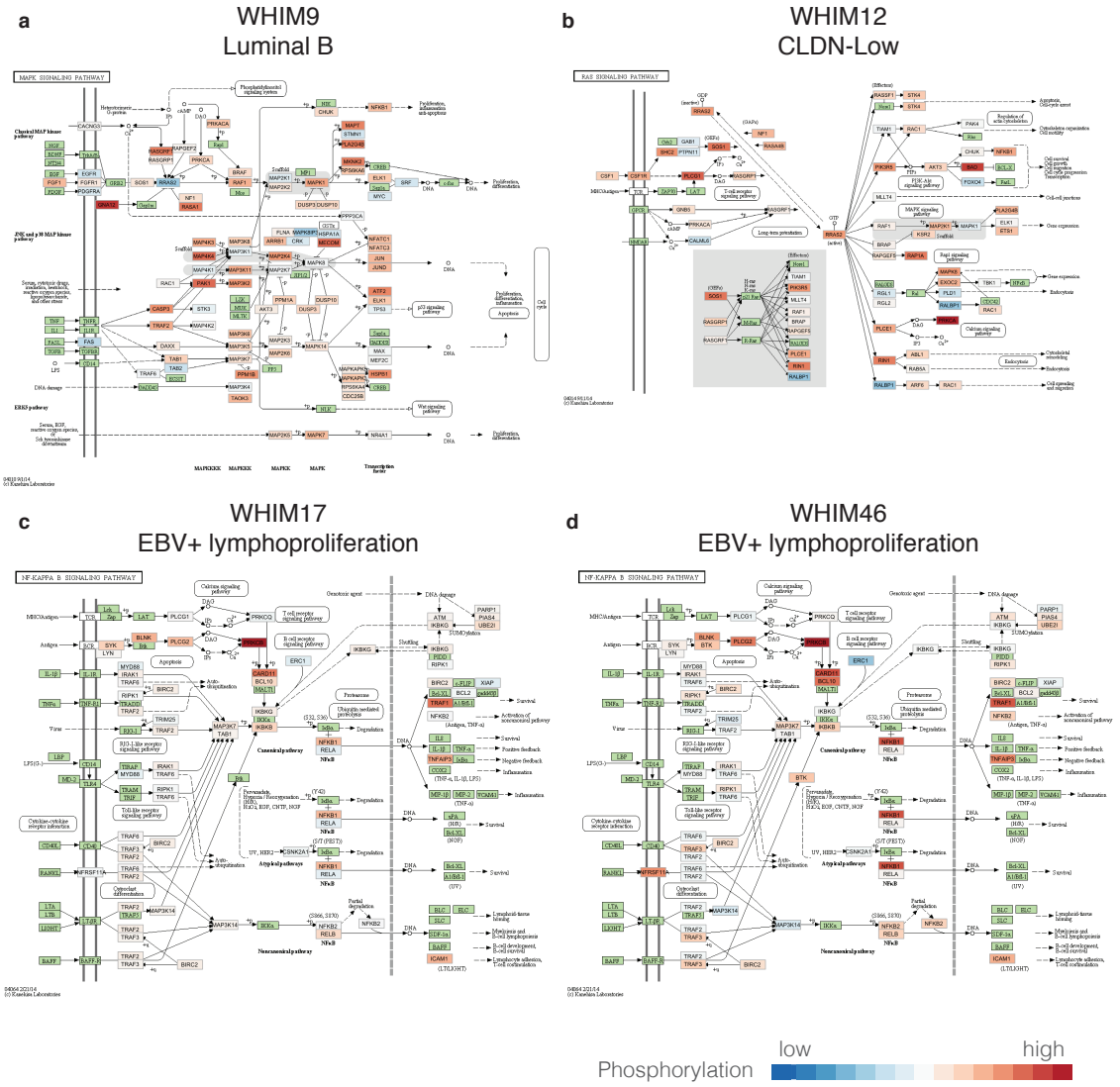


Figure 3.4. Activated signaling pathways detected through pathway phosphorylation enrichment analysis. (a) Activation of the MAPK signaling pathway in WHIM9. (b) Activation of the Ras signaling pathway in WHIM12. Phosphorylation levels of each protein in the pathway relative to the cohort of 24 PDX models are shown by the color scale of red (high) to blue (low). Proteins with no phosphorylation data are colored in green.

While pathways such as PI3K/AKT/MTOR are known to be activated across the majority of breast tumors[88], our analysis shows that other complementary tumorigenesis-related pathways including RAS/MAPK are also activated in a small set of breast tumors due to specific genomic or proteomic alterations, representing alternative treatment opportunities. In addition, our phospho-proteomic analysis revealed activation of signaling pathways not readily predicted by genomic data.

Complementary Genomic/Proteomic Druggable Targets

We examined promising drug targets in each tumor by surveying the genome and expressed proteomes. Specifically, we compiled a list 76 druggable genes, along with their respective drugs, from established public databases (Methods). Six PDXs, representing 20.8% of the tumors in this study, harbored druggable somatic mutations, including *PIK3CA* p.H1047R and *KRAS* p.A146V in WHIM9, *PIK3CA* p.H1047R in WHIM16 and WHIM24, *PIK3CA* p.E545K in WHIM18, *PIK3CA* p.E542K in WHIM20, and *SF3B1* p.K700E in WHIM26.

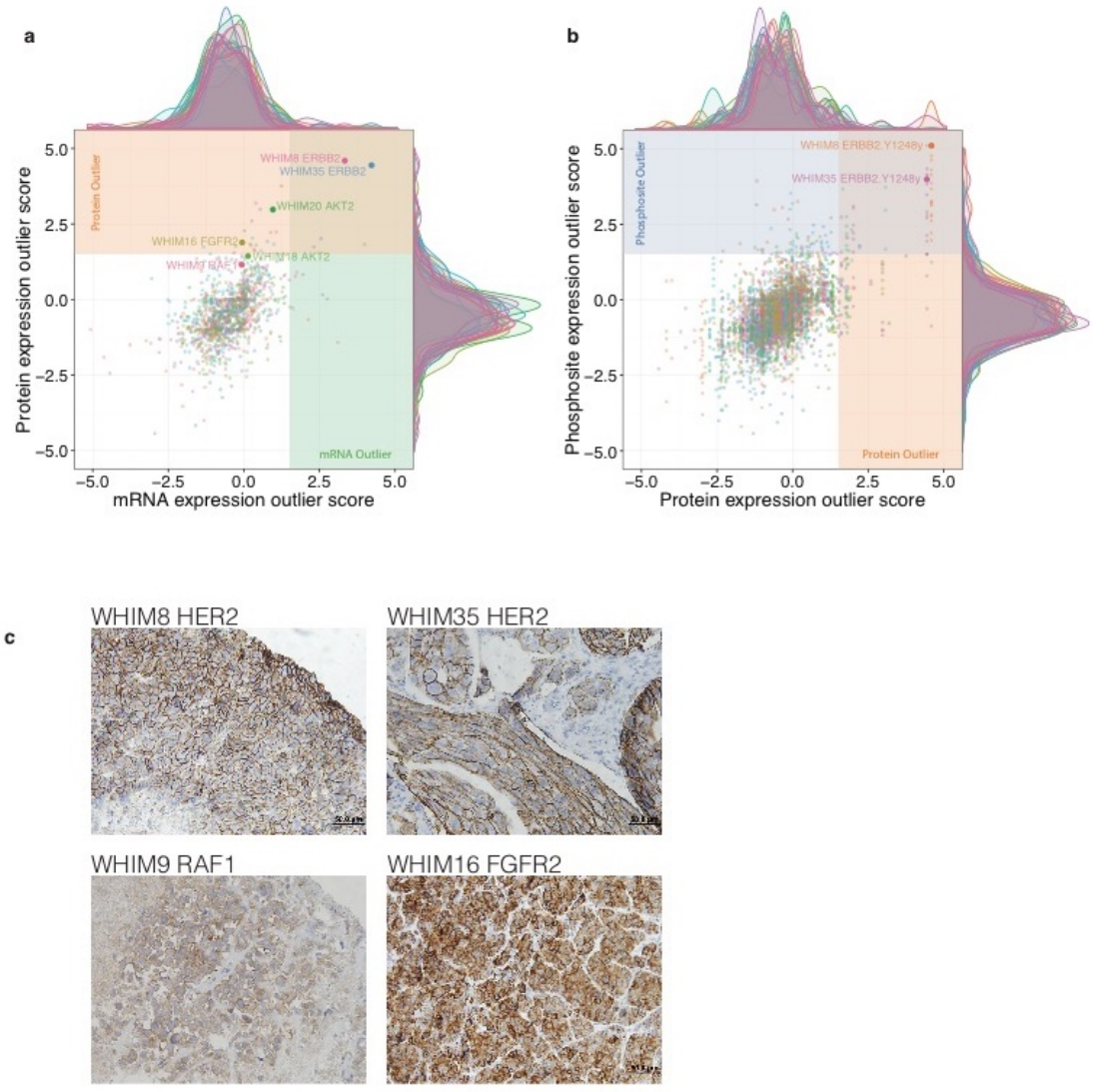


Figure 3.5. Outlier expression analysis identified druggable events at mRNA expression, protein expression, and protein phosphorylation levels in 24 breast cancer PDXs. Druggable outlier events identified at (a) the mRNA and protein and (b) protein and phosphopeptide expression levels. Each color represents one xenograft sample. Key outlier events validated in this study are labeled by text. (c) Immunochemistry staining verified outlier expression of HER2 in WHIM8 and WHIM35, RAF1 in WHIM9, and FGFR2 in

WHIM16.

While activating mutations in oncogenes can be targeted by treatments, aberrantly over-expressed or activated protein products, such as HER2, also presents exploitable treatment opportunities[91-93]. We sought genomic and proteomic evidence of over-expressed genes/proteins or proteins with highly phosphorylated sites. We defined outliers as expression values exceeding the 1.5 interquartile ranges (IQR) above the third quartile of the cohort (Variations of Box Plots), and rank ordered them by the outlier score. We further required CNV outliers to be validated with outlier expression in either the mRNA or protein level to rule out up-regulation events due to technical artifacts or passenger events (Methods). mRNA and protein expression outlier scores showed a moderate positive correlation (Figure 3.5a, R = 0.516), but mRNA outlier expression did not guarantee high protein expression (e.g., AKT2 and FGFR2, Figure 3.5a). Similarly, we observed a fraction of phosphosite outliers not detected at the protein level (Figure 3.5b, R = 0.548). Consequently, identifying post-transcriptional and post-translational events to capture potential druggable treatment opportunities requires consideration of protein expression as well as gene expression.

Applying this druggable outlier detection strategy across CNV, mRNA, protein, and protein phosphorylation levels, we identified over-expressed druggable genes in 26.1% and 47.8% of PDXs at the CNV and mRNA levels, respectively (Figure 3.6). These events recapitulated known druggable opportunities, such as the *PIK3CA* copy number amplification in WHIM4 and *ERBB2* (*HER2*) copy number amplification in WHIM35. Expanding to iTRAQ protein expression outliers allowed us to uncover druggable targets in 19 out of the 24 PDXs (79.2%),

while considering phosphosite outliers covered 22 of the 24 PDXs (91.7%). A significantly high fraction of protein outliers overlapped between the LFQ and iTRAQ datasets (Fisher's Exact Test, P = 2.013e-05), providing validation of our findings.

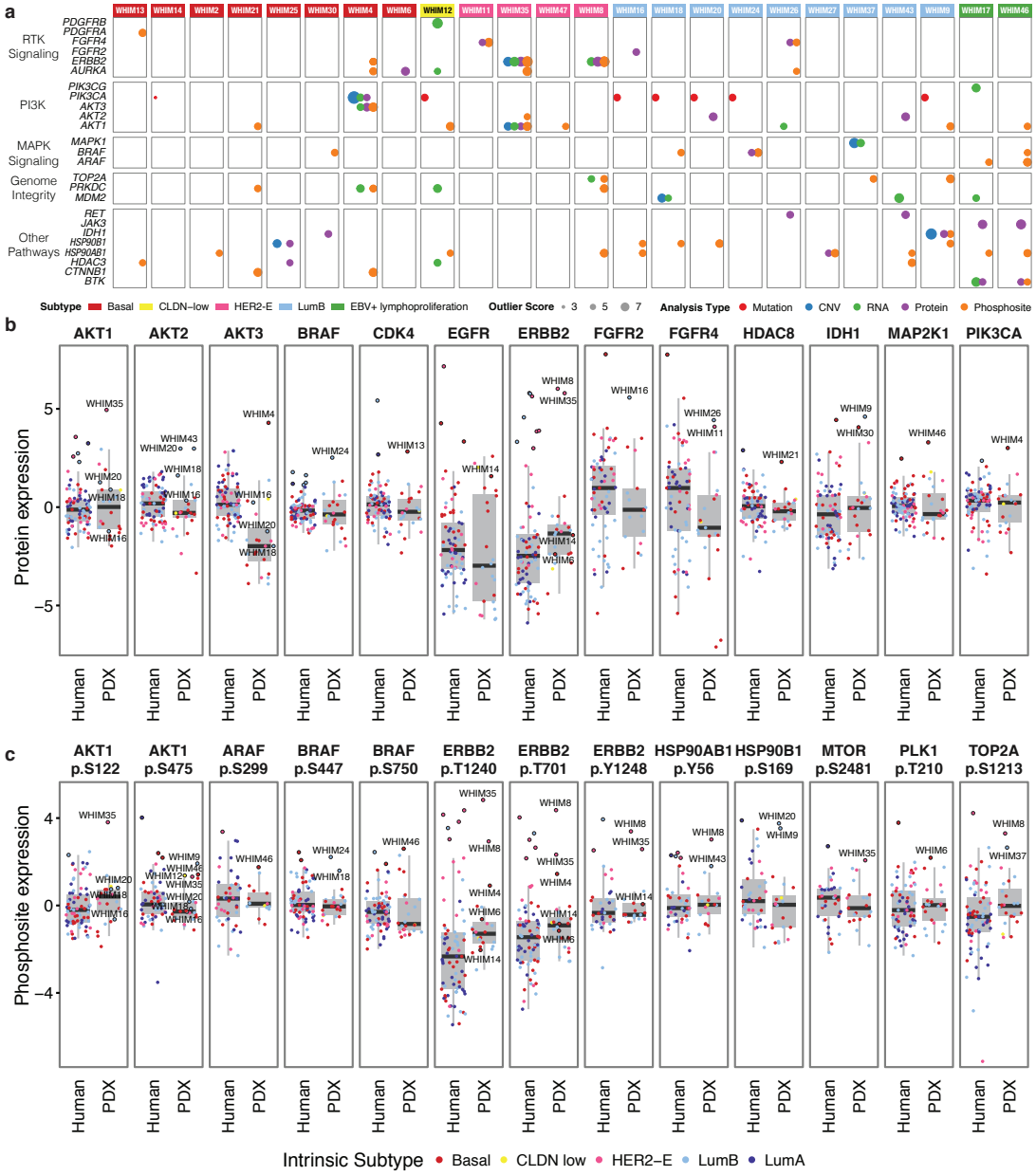


Figure 3.6. Druggable outlier events identified at CNV, RNA, protein, and phosphosite

levels of 24 PDX models and 77 TCGA human breast tumors. (a) Outlier analysis revealed potentially druggable events in the RTK, PI3K, MAPK signaling, genome int pathways at various frequency and magnitudes across 4 breast cancer subtypes. Selecte genes with any outlier score greater than 2.5 or in the key oncogenic pathways, includi PI3K, RTK, MAPK signaling pathways, are shown. (b) Comparison of protein express outliers of selected druggable genes in PDX and human breast tumors. (c) Comparison over-expressed phosphosite outliers of selected druggable genes in PDX and human br tumors. Key outlier events reaching the outlier definition threshold or validated in this are labeled by text.

The identified proteomic outliers included multiple known druggable targets involved in the PI3K, RTK, MAPK signaling pathways and other oncogenic processes (Figure 3.6). For instance, HER2 was found to be the top outlier at the protein expression level in 2 HER2-E xenografts, WHIM8 and WHIM35; the phosphosites of HER2 were also identified as outliers. Immunohistochemistry experiments validated the high HER2 protein expression on the cell membranes of WHIM8 and WHIM35 (Figure 3.5c). Further, AKT1, AKT2 and AKT3 proteins were outlier expression candidates in WHIM35, WHIM20 and WHIM43, and WHIM4, respectively, providing proteomic validation of previously observed AKT up-regulation at mRNA expression level in breast cancer[88]. Other outlier proteins included IDH1 in WHIM9 and WHIM30, FGFR4 in WHIM11 and WHIM26, and both BTK and JAK3 in the EBV-positive WHIM17 and WHIM46. Many of these protein outliers aligned with up-regulation in their corresponding phosphosites, including phosphosites on AKT1, AKT3, BRAF, FGFR4 and HSP90AB1 (Figure 3.5c). We also discovered additional outlier phosphosites in other proteins including CTNNB1, ARAF, and HSP90B1. Finally, we identified outlier FGFR2 in WHIM16 and high RAF1 in WHIM9 and further validated their high protein expression status of by immunohistochemistry analyses (Figure 3.5c). As diverse sets of

proteomic outliers are identified across PDXs, effectively inhibiting such activation events will be required when designing targeted treatment strategies to each individual breast tumor.

Notably, for 80.6% of proteomic outlier events, we were able to identify human tumors from the 77 TCGA samples[83] showing the same outlier protein expression through outlier score or ranking (Methods, Figure 3.6b). For example, we observed outlier HER2 expression in 5 HER2-E and 4 luminal B human breast tumors. Further, a basal human sample carried both the outlier FGFR2 and FGFR4 expression, validating the findings in WHIM16, WHIM26 and WHIM11. Basal human breast tumors also carried protein expression outliers in IDH1, EGFR and MAP2K1 (Figure 3.6b). Phosphosite outlier expression events showed a moderate rate of validation in the same human cohort (48.8%, Figure 3.6c), suggesting its transient nature and potential micro-environmental effects on protein phosphorylation. As expected, HER2-E and a few HER2-positive luminal B tumors carried outliers in HER2 phosphosites including p.T701, p.T1240 and p.Y1248. ARAF phosphosite outliers, such as p.S299, were found in both human and PDX samples, validating our previous finding in the human cohort[12]. Interestingly, we identified outlier phosphosites in genes not previously implicated in breast cancer through genomic profiles, such as BRAF p.S447, p.S750 and HSP90AB1 p.Y56 and p.S169 (Figure 3.6c). Our results demonstrated that proteomic outlier events, like genomic driver mutations, are consistently observed in PDXs and human tumors. Some protein outlier events might represent "proteomic drivers" of tumorigenesis and therefore potential drug targets in breast tumors.

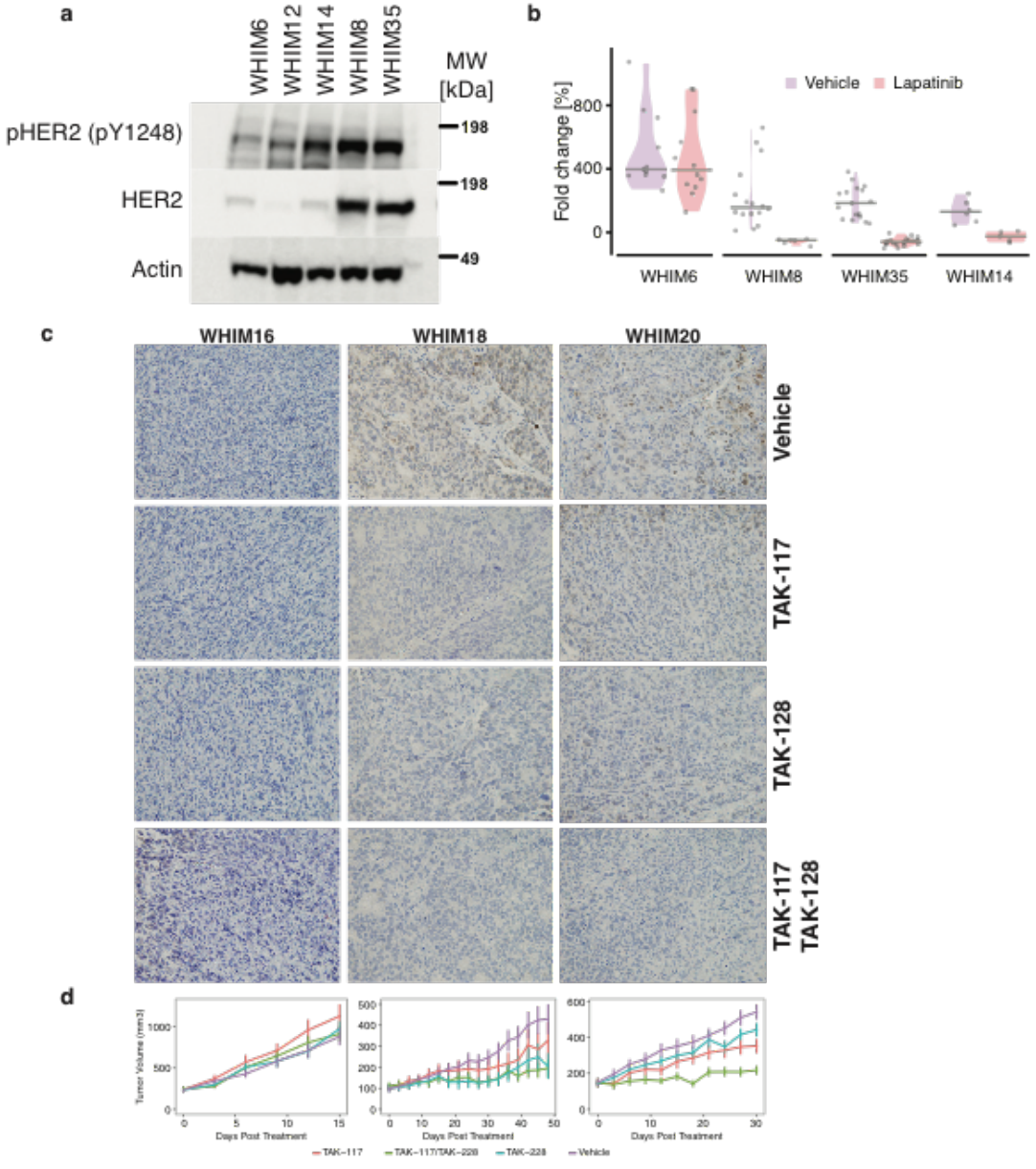


Figure 3.7. Targeted treatments of breast cancer xenografts. (a) Western blot of HER2 protein and HER2 p.Y1248 expression levels in 5 WHIM models (WHIM6, WHIM8, WHIM12, WHIM14, and WHIM35). (b) *In vivo* treatment responses to lapatinib in 4 PDX

models including two HER2 positive lines (WHIM8 and WHIM35) and two basal line
(WHIM6 and WHIM14). The response is measured in fold change (%) of tumor volun
after 2 weeks of vehicle or lapatinib treatment. (c) Immunochemistry staining of AKT
phosphorylation status in WHIM16, WHIM18, and WHIM20 in response to PI3K inhi
TAK-117, mTOR inhibitor TAK128, and TAK-117/TAK-128 combined. (d) *In vivo*
treatment responses to PI3K/mTOR inhibitors (TAK-117/TAK-128) of WHIM16, WH
and WHIM20. Values were represented by tumor volume [mm^3] every 3 day followin
treatment.

To validate the identified druggable events, we conducted treatment experiments targeting HER2 and PI3K pathways on selected PDX models. Four PDX models were chosen to address HER2 targeting using lapatinib, an oral HER2 kinase inhibitor. These include 2 HER2-E PDX models WHIM8 and WHIM35, both with high HER2 protein and phosphoprotein expression, and 2 low HER2 expressing, basal-like PDX models WHIM6 and WHIM14. Western blotting suggested high HER2 expression and phosphorylation of HER2 p.Y1248 in both HER2 positive tumors, WHIM8 and WHIM35 (Figure 3.7a). Unexpectedly, high levels of HER2 p.Y1248 phosphorylation was also detected in WHIM14 by Western blotting. In contrast MS-based phosphoproteomics also detected high levels of HER2p.Y1248 in WHIM8 and WHIM35, but not in WHIM14, which was more consistent with the known biology of the models. Antibody-based diagnosis of HER2 activation in WHIM14 was possibly due to cross-reaction of the Y1248 antibody with the pY1172 site of EGFR that bears high sequence similarity around the pY residue (..GTPTAENPEy_{1248}LGLDVPV-CO_2H vs. ..GSTAENAEy_{1172}LRVAPQ..). As expected, WHIM8 and WHIM35 were growth inhibited (Wilcoxon Rank Sum Test, P = 2.7e-5, 4.8e-7) by lapatinib, whereas WHIM6 was not (P = 0.65) (Figure 3.7b). Interestingly, WHIM14 also showed significant reduction in tumor growth by lapatinib (P = 4.3e-3). Upon further exploration

with a lower, clinically achievable dose of lapatinib for chronic treatment (30mg/kg compared to 220mg/kg in the previous experiment) over 48 days, significant but weak tumor growth inhibition was again achieved (P = 0.0311). However, WHIM14 did not respond to HER2/HER3 antibodies trastuzumab or pertuzumab (P = 0.250 and 0.181, respectively). Thus, the response of WHIM14 to lapatinib was likely due to inhibition of EGFR not HER2[94,95]. While WHIM6 also showed elevated EGFR protein level, WHIM14 showed notably higher EGFR phosphorylation based on both mass spectrometry (Figure 3.2b) and western blotting (Figure 3.7a), which could account for their different response to lapatinib. Two basal and one HER2-E human breast cancers harbored outlier EGFR expression (Fig 6b), suggesting EGFR remains a potential therapeutic target in a subset of breast cancers that has yet to be fully realized clinically.

The PI3K/AKT/mTOR signaling pathway is altered in approximately 77% of breast tumors[88] and multiple drugs targeting its components, including Class I PI3Ks, AKTs, and mTORs, are already in clinical trials[96]. Among these, a combination of everolimus and exemestane has been approved for treating advanced ER-positive breast cancers resistant to non-steroidal aromatase inhibitors[97]. Promising activity has also been reported for direct inhibitors of PI3K[98,99]. However, mutations in PIK3CA and other genetic alterations at the genomic level have so far failed to closely predict therapeutic responsiveness to PI3K pathway inhibitors[98]. We therefore hypothesized that combined genomic and proteomic indication of PI3K signaling activation is necessary for the prediction of treatment response. Our integrated approach identified 6 xenografts that harbored complementary genomic and proteomic druggable events in the PI3K-AKT pathway. In particular, WHIM16 harbored a hotspot *PIK3CA* p.H1047R mutation, whereas WHIM18 and WHIM20 each carried a hotspot *PIK3CA* p.E545K mutation. WHIM20 also

showed an additional outlier protein expression of AKT2 and WHIM18 also expressed AKT2 at a near outlier level that may combine to activate PI3K pathway signaling (Figure 3.6). Since the treatments were applied to later-passage PDX models relative to the ones we conducted proteogenomic analysis on, we performed immunohistochemistry to validate the phosphorylation of AKT p.S473. WHIM18 and WHIM20 showed detectable AKT p.S473, which was not observed in WHIM16 (Figure 3.7c).

We conducted combinatorial treatment experiments by applying an alpha specific PI3K inhibitor (TAK-117) and/or an mTORC1/2 inhibitor (TAK-228) to three PDX models of luminal B breast cancer. Consistent with previous reports, *PIK3CA* mutation status alone did not accurately predict outcome; WHIM18 and WHIM20 showed reduced tumor growth upon application of the inhibitors, whereas WHIM16 did not (Figure 3.7d). mTOR inhibition repressed tumor growth in WHIM18 and WHIM20 (ANOVA followed by Tukey's post hoc test, $p = 9.5e\text{-}10$, $1.6e\text{-}03$ respectively), but not in WHIM16 ($p = 0.97$), showing that inhibition of mTORC1/2 may effectively suppress breast tumors with activated AKTs and validating a previous study showing the efficacy of mTOR inhibition in PDX models of triple-negative breast cancer[100]. Importantly, the combinatorial treatment achieved the greatest effect in WHIM18 and WHIM20 ($p < 2.2e\text{-}16$ for both comparisons). While neither PI3K nor mTOR inhibitor drug treatment alone suppressed tumor growth in WHIM20 completely, the combination of both PI3K and mTOR inhibitors significantly reduced tumor growth to a nearly static state (Figure 3.7d). Based on our proteomic characterization, WHIM20 exhibited the strongest AKT1 and AKT2 protein expression, as well as AKT1 p.S122 and p.S475 phosphorylation signatures followed by WHIM18 and then WHIM16 (Figure 3.6b, c). Further, both PI3K and mTOR inhibitors significantly reduced AKT

p.S473 (Figure 3.7c). Our treatment results showed that the magnitude of drug response may be associated with over-expression and phosphorylation of the downstream signaling targets such as AKT proteins.

In addition to this validation of druggable hypotheses in luminal tumors, our previous report also demonstrated the effectiveness of combinatorial therapy of AKT and mTOR inhibitors in two other basal breast cancer xenografts[101]. One of the treated xenografts, WHIM4, was also characterized in this study, and showed copy number amplification of *PIK3CA* and AKT3 protein outlier expression (Figure 3.6). Overall, proteogenomic analysis revealed that the dual activation of PIK3CA at the genomic level and AKTs at the protein level may be a common signature of breast tumors, affecting more than 20% of PDXs in this cohort. Importantly, our results demonstrate the potential utility of combinatorial inhibitor treatments to treat breast tumors showing these proteogenomic signatures.

3.3 Discussion

Breast cancer has been traditionally characterized in the clinic through hormone receptor status and selected genes' expressions[86,102], and more recently by genomic sequencing[13]. However, druggable genomic driver events are detectable in only a limited percentage of patients[13]. As the majority of drugs target proteins, a systematic evaluation of breast cancer proteomes would seem ultimately to be necessary for selecting targeted treatment and predicting drug response. Recent

advances in MS-based proteomics allow extensive and quantitative surveys of the global proteome. Here, we have systematically analyzed proteogenomic profiles of 22 patient-derived breast cancer xenografts and 2 EBV-positive lymphoproliferations that are likely artifacts of engraftment of human lymphocytes into NSG mice.

This study shows that proteogenomic signatures of PDXs resemble most findings from breast cancer patients. While some discrepancies exist, we established a normalization strategy at both the genomic and proteomic levels that enabled direct comparison. PDX tumors recapitulated the proteomic diversity of human breast cancers (Figure 3.3). We also identified multiple druggable targets for each tumor model (Figure 3.6). Proteomic events validated a significant number of CNV or mRNA up-regulations. For example, HER2 protein and phosphosite outlier expression was observed in HER2-E WHIM8 and WHIM35 (Figure 3.5), which were effectively treated using lapatinib (Figure 3.7b). Interestingly, we also identified overexpressed proteomic events not evident in the genomic level in both PDX and human samples, including outlier protein expression of EGFR, and outlier phosphosite expressions of ARAF, BRAF, HSP90AB1, PTPN11 and TOP2A (Figure 3.6), highlighting potential new treatment opportunities in breast cancer. In the two PDX models subsequently diagnosed as EBV-positive lymphoproliferations, we observed outlier BTK expression (Figure 3.6) and activation of the NFκB pathway (Figure 3.4), validating BTK as a druggable target in EBV+ lymphomas[103]. While more than 80% of the proteomic outlier events in PDX were also found in human tumors, a lessor 48.8% of phosphosite outlier events were validated, potentially due to different tumor micro-environments. Thus, transient phosphoproteomic events identified in PDX tumors would likely require further verification in their corresponding primary tumors.

Outlier protein expression events can likely lead to downstream pathway activation, such as MET outlier protein expression (Figure 3.6) and activated Ras pathway observed in WHIM12 (Figure 3.4). While genomic analysis has utilized mutual occurrence or exclusivity in patient cohorts to deduce pathway relationships, phosphorylation profile analyses allowed us to directly interrogate signaling in these established pathways in a single sample. Our pathway activation results suggest these events may be crucial to tumorigenesis and that some are likely the proteomic "smoking guns" that originally triggered the oncogeneic cascade.

Roughly 77% of breast tumors showed alterations in the PI3K-AKT signaling pathway, representing a potentially important path forward for drug-based treatment. Yet, genomic alterations of PI3K pathway components have not shown themselves to predict treatment responsiveness to PI3K pathway inhibitors[98]. In this study, we defined complementary druggable targets of the pathway using proteogenomic analysis, including several events of a co-occurring PIK3CA mutation or copy number amplification, and AKT protein outlier expression coupled by elevated AKT phosphorylation. In two such breast tumors, we successfully inhibited tumor growth using a combination of PI3K and MTOR inhibitors (Figure 3.7c). While these results show potential functional implications, additional, systematic treatment experiments are required to validate the identified proteomic druggable targets.

In conclusion, this initial work using proteogenomic integration coupled with patient-derived xenograft validation, has demonstrated a strategy that, in principle, may enable more accurate prediction of the efficacy of mechanism-based cancer therapeutics.

3.4 Methods

Xenograft Model Generation

Patient-derived xenografts were generated from primary or metastatic breast tumors using previously described procedures[19]. All human tissues for these experiments were processed in compliance with NIH regulations and institutional guidelines, and approved by the institutional review board at Washington University. All animal procedures were reviewed and approved by the institutional animal care and use committee at Washington University in St. Louis. PDX models are available through the application to the Human and Mouse-Linked Evaluation of Tumors core at http://digitalcommons.wustl.edu/hamlet/.

We selected 24 of the established breast tumor samples, including 9 luminal B, 10 basal, 4 HER2-E, and 1 CLDN-low breast tumors for further proteogenomic characterization. Receptor statuses of xenograft tumors were validated using IHC after engraftment.

Immunohistochemistry

Xenografts were formalin-fixed at least for 24 hours and paraffin-embedded. Sections were evaluated by hematoxylin & eosin (H&E) staining. Immunohistochemistry (IHC) was performed on additional sections for HER2 (Dako), Phospho-Akt (Ser473) (Cell signaling), FGFR2 (Abcam), and Raf-1 (Santa Cruz) following the manufacturer's instructions. Western blotting of phosphorylated HER2 (p.Y1248) was performed using antibody cat Nr-06-229 (Millipore).

In-vivo Drug Treatment Experiments

For the targeted treatment of the PI3K pathway, PI3K alpha inhibitor TAK-117 and TORC1/2 inhibitor TAK-228 were provided by Millennium Pharmaceuticals, Inc. The compounds were dissolved in Peg400. Tumors were engrafted in NSG (NOD.Cg-*Prkdc*^{scid} *Il2rg*^{tm1Wjl}/SzJ) mice (The Jackson Laboratory) by subcutaneous injection of 2-5 × 10^6 PDX cells in PBS supplemented with 30% Matrigel (BD Biosciences Cat. No. 354234). When tumors reached an average size of 250-300 mm^3, animals were assigned randomly to control and various treatment groups (n=8-9 each group). For treatment of WHIM16 (passage 7), 18 (passage 8), and 20 (passage 5) were used. Tumor bearing mice were gavaged with: 1) Peg400; 2) TAK-117, 140 mg/kg/day; 3) TAK-228, 1mg/kg/day; 4) TAK-117, 140 mg/kg/day, TAK-228, 1mg/kg/day. The mice were treated on three consecutive days once daily and then had a 4-day interval. Tumors were measured with external caliper, and volume was calculated as $(4\pi/3) \times (\text{width}/2)^2 \times (\text{length}/2)$.

For the lapatinib therapeutic experiments of WHIM6 (passage 7), WHIM 14 (passage 11), WHIM 8 (passage 6) and WHIM 35 (passage 6), 1x10^6 tumor cells were added to equal volume of 1:1 mixture of Matrigel and 10% RPMI plus 10% FBS to 4th mammary fat pads in female SCID/bg mice (ENVIGO). We then established tumors to an average volume 250-300mm^3, and randomized the mice into control and lapatinib groups. We treated the treated group mice with lapatinib chow diet by formulating lapatinib (220mg/kg) and processing them into food pellets (by Research Diets Inc), which is supplied for 2 weeks. The numbers of replicates for each group are as followed: WHIM 14 control (n=6) and lapatinib treated (n=5); WHIM 6 control (n=11) and lapatinib treated (n=12); WHIM 8 control (n=16) and lapatinib treated (n=6); WHIM 35 control (n=17) and lapatinib treated (n=18). We also treated WHIM14 with 100mg/kg lapatinib

treatment for 48 days, including both control (n=8) and lapatinib treated (n=9) groups. Further, we tested the effect of trastuzumab (30mg/kg weekly with IP injections) and pertuzumab (30mg/kg weekly with IP injections) on PDX tumor growth for 48 days. The experimental groups are as following: control group (n=6) treated with physiological saline (vehicle); trastuzumab treated group (n=12); pertuzumab treated group (n=11).

Statistical testing of the resulting data was conducted using the R programming language. Wilcoxon rank sum test was applied to compare the fold change in tumor volumes after two weeks in the lapatinib treatment experiment, and after 44 days or 48 days for the additional WHIM14 experiment using low-dose lapatinib and trastuzumab/pertuzumab. One-way ANOVA was applied to compare the treated vs. control groups in the PI3K targeted-therapy experiments, and a follow-up Tukey's post-hoc test was used for the various comparisons in PI3K inhibition experiments. All human tissues for these experiments were processed in compliance with NIH regulations and institutional guidelines, and approved by the Washington University, University of North Carolina, or Baylor College of Medicine Institutional Review Board (IRB). All animal procedures were reviewed and approved by the institutional animal care and use committee at Washington University in St. Louis, University of North Carolina, or Baylor College of Medicine.

Genomic and Proteomic Data Generation

Somatic Mutation

Sequencing reads were aligned using BWA[104]. Somatic variants were identified using VarScan2[105-107], GATK[108], and Pindel[109], and annotated based on Ensembl release 70_37. We

then filtered out common variants by using variants from the 1000 Genomes and NHLBI projects. We further eliminated mouse contamination by filtering somatic variants that were mapped to mouse reference genome. Somatic mutation calls were validated either using a custom array or manually reviewed in IGV. The genomic data of 17 out of the 24 PDXs have been presented before in our previous studies[16,110].

Copy Number Variation

The segment-based copy number data were generated using the whole-genome sequencing and exome sequencing data. We then converted the segment-based copy number data to the gene-based copy number data by using the RefSeq database (version 20130727). The copy number values were further transformed to the log-R ratio, using the cohort mean for the gene as the reference.

mRNA Expression and virus detection through RNA-Seq

mRNA expression values were calculated from mRNA sequencing data using MapSplice[111,112]. The resulting RSEM values were normalized within samples to a fixed upper quartile. Upper quartile normalized RSEM data were log2 transformed and the data were median centered by gene. To quantify virus abundance, we used the VirusScan pipeline (https://github.com/ding-lab/VirusScan) to detect viruses by numbers of virus-supporting reads from RNA-Seq data.

LFQ proteome

Tumor Sample Generation and Protein Extraction: Patient-derived xenograft breast tumors were processed to cryopulverized powders as described previously[113]. The powders (100 mg wet weight) were subjected to lysis and protein extraction using a buffer composed of 8 M urea, 50

mM Tris pH 8.0, 75 mM NaCl, 1 mM $MgCl_2$, and 500 units Benzonase. Approximately 1 mg of total protein extracted was reduced with DTT and subsequently alkylated with iodoacetamide. The proteins were then subjected to proteolysis with endoproteinase Lys-C (Wako Chemicals, USA) for ~4 hours at 37^0C. The solution was diluted 4-fold with 25 mM Tris pH 8.0, 1 mM $CaCl_2$ and further digested with trypsin (Promega) for ~12 hours at 37 ^{0}C. Digestion was stopped by the addition of TFA to 0.4%, and the precipitate was removed by centrifugation. The peptide solutions were desalted on Sep-Pak Light C18 cartridges (Waters) and dissolved in 30% ACN, 0.1% TFA before loading on a 300 μm Source 15S (GE Healthcare) column for Basic Reversed Phase Chromatography (bRPLC)[49]. A linear LC gradient was performed by increasing buffer B from 0-70% within 60 min, where buffer A was aqueous 10 mM ammonium formate, and buffer B was 90% ACN in 10 mM ammonium formate. A total of 30 fractions were collected for each WHIM sample (18 WHIMs). Five fractions were then prepared by combining non-contiguously fractions. We analyzed an additional technical replicate for WHIM2 and WHIM16. The fractions were dried and desalted using a stop-and-go-extraction tip (StageTip) protocol containing 4 x 1 mm C18 extraction disk (3M).

Liquid Chromatography-Tandem Mass Spectrometry and Protein Identification:

Sample analysis was performed via reversed phase LC-MS/MS using a Proxeon 1000 nano LC system coupled to a Q Exactive mass spectrometer (Thermo Scientific, San Jose, CA). The Proxeon system was configured to trap peptides using a C18 column (3 cm x 100 μm i.d.) with a diverted flow rate (5 μL/min) The trap column was placed in line with the analytical column (15 cm x 75 μm i.d., 3.5 μm, 300 Å particle C18, Thermo Scientific, San Jose, CA) prior to gradient

elution of peptides. Analytical separation of all the tryptic peptides was achieved with a linear gradient of 2-30% buffer B over 240 min (250 nL/min), where buffer A was aqueous 0.1% formic acid, and buffer B was acetonitrile in 0.1% formic acid.

LC-MS experiments were performed in a data-dependent mode with Full-MS (externally calibrated to a mass accuracy of < 5 ppm, and a resolution of 70,000 at m/z = 200) followed by HCD-MS/MS of the top 20 most intense ions. High-energy collision activated dissociation (HCD)-MS/MS was used to fragment peptides at a normalized collision energy of 27 eV in the presence of nitrogen. One LC-MS run was performed for each fraction (from 1 process technical replicate), except for WHIM2 and WHIM16 where 2 LC-MS runs were conducted (from 3 process technical replicates), resulting in the production of 100 LC-MS runs for global peptide analysis. Mass spectra were processed, and peptide identification was performed using the Andromeda search engine found in MaxQuant software ver. 1.5.0.25. (Max Planck Institute, Germany). All protein database searches were performed against the RefSeq database (version 20140707). Peptides were identified with a target-decoy approach using a combined database consisting of reverse protein sequences of the RefSeq human, mouse and common repository of adventitious proteins (cRAP). The cRAP database was obtained from the Global Proteome Machine (**ftp://ftp.thegpm.org/fasta/cRAP**). Peptide inference was made with a false discovery rate (FDR) of 1% while peptides were assigned to proteins with a protein FDR of 5%. A precursor ion mass tolerance of 20 ppm was used for the first search that allowed for m/z retention time recalibration of precursor ions that were then subjected to a main search using a precursor ion mass tolerance of 6 ppm and a product ion mass tolerance 0.5 Da. Search parameters included up to two missed cleavages at KR on the sequence, and oxidation of methionine, and protein *N*-terminus acetylation as a dynamic modification.

Carbamidomethylation of cysteine residues was considered as a static modification. Peptide identifications are reported by filtering of reverse and contaminant entries and assigning to their leading razor protein according to the Occams razor principal. The mass spectrometric data are deposited at the CPTAC Data Coordinating Center as raw and mzML files (**https://cptac-data-portal.georgetown.edu**)[114].

Peptide and Protein Quantitation: Label-free quantitation (LFQ) was performed based on peak areas. The measured area under the curve of *m/z* and the retention time-aligned extracted ion chromatograms of peptides were performed via the label-free quantitation module in MaxQuant [ver. 1.5.0.25][115]. All replicates for each PDX were included in the LFQ experimental design with peptide-level quantitation performed using unique and razor peptide features corresponding to identifications filtered with a posterior error probability (PEP) of 0.01, peptide FDR of 0.01 and protein FDR of 0.05. The expression values were median centered in the Perseus software for further analysis [version 1.5.0.9].

<u>iTRAQ proteome and phosphoproteome</u>

We included all 24 of the established breast tumor samples for proteomic characterization using iTRAQ. Tumor tissue samples were maintained in cryovials at -80°C until cryopulverization using a CP02 Cryprep Pulverizer (Covaris, Woburn, MA). 90 mg aliquots of cryofractured material were prepared for proteomic processing in aluminum weighing dishes on dry ice using spatulas kept cold in liquid nitrogen, with remaining material reserved for other applications. The 90 mg target was designed to include 40 mg for each of the collaborating research teams, with an anticipated yield for each team of 1.5 – 2 mg protein based on 4-5% recovery. To avoid

systematic bias, sample processing was block randomized, with each intrinsic subtype proportionally represented in each processing tranche.

The reproducibility of the iTRAQ4-plex global proteome and phosphoproteome analysis workflow used in this study has been extensively tested for quantitative reproducibility both within and across laboratories in the CPTAC program[83,116]. Over a period of several months 5 iTRAQ4-plex replicates were measured at each of the 3 CPTAC proteome analysis centers. Each of these iTRAQ4-plexes contained duplicate measurements for both a basal WHIM2 and a luminal WHIM16 PDX samples that are also part of this study. A high degree of consistency in the number of proteins identified and correlation in the protein expression was obtained[116]. Pearson correlations for replicate proteome and phosphoproteome measurements were very high with a r=0.9 in our previous study[83] and very similar to the correlation observed here for the WHIM13 replicate measurement. These data show that our platform provides highly reproducible quantitative measurements for global proteomes and phosphoproteomes.

Protein extraction, digestion and iTRAQ labeling of peptides from breast cancer tumors:

Cryopulverized breast cancer tumor samples tissues (~2 combined aliquots of 90 mg tissue weight each) were homogenized in 1000 µL lysis buffer containing 8M urea, 75mM NaCl, 1mM EDTA in 50mM Tris HCl (pH 8), 10 mM NaF, phosphatase inhibitor cocktail 2 (1:100; Sigma, P5726) and cocktail 3 (1:100; Sigma, P0044), 2 µg/mL aprotinin (Sigma, A6103), 10 µg/mL Leupeptin (Roche, #11017101001), and 1 mM PMSF (Sigma, 78830). Lysates were centrifuged at 20,000 g for 10 minutes before measuring protein concentration of the clarified lysates by BCA assay (Pierce). Protein lysates were subsequently reduced with 5 mM dithiothreitol (Thermo Scientific, 20291) for 45 minutes at room temperature, and alkylated with 10 mM

iodoacetamide (Sigma, A3221) for 45 minutes in the dark. Samples were diluted 4-fold with 50mM Tris HCl (pH 8) prior to digesting them with LysC (Wako, 129-02541) for 4 hours and trypsin (Promega, V511X) overnight at a 1:50 enzyme-to-protein ratio at room temperature overnight on a shaker.

Digested samples were acidified with formic acid (FA; Fluka, 56302) to a final volumetric concentration of 1 % or final pH of ~3-5, and centrifuged at 2,000 g for 5 minutes to clear precipitated urea from peptide lysates. Samples were desalted on C18 SepPak columns (Waters, 100mg, WAT036820) and 1mg peptide aliquots were dried down using a SpeedVac apparatus.

Construction of the Common internal Reference Pool: The proteomic and phosphoproteomic analyses of xenograft samples were performed as iTRAQ 4-plex experiments. Quantitative comparison between all samples analyzed was facilitated by the use of iTRAQ reporter ion ratios between each individual sample and a common internal reference sample present in each 4-plex. The reference sample was comprised of 16 of the 24 WHIM tumors analyzed in this study with equal contribution for each tumor (WHIM numbers 2, 4, 6, 8, 11, 12, 13, 14, 16, 18, 20, 21, 24, 25, 30, and 46). The 24 tumor samples were analyzed in 9 independent 4-plex experiments, with 3 individual samples occupying the first 3 channels of each experiment and the 4th channel being reserved for the reference sample. While 8 iTRAQ 4-plex experiments were used to analyze the 24 individual WHIM tumor samples, an additional 4-plex experiment was designed to include the WHIM13 sample for process replicate analysis and also internal reference samples from our human primary breast cancer study[83] and a taxol drug response study (unpublished) to allow cross-referencing of the different datasets.

iTRAQ labeling, high pH reversed-phase separation and phosphopeptide enrichment of peptide samples: Desalted peptides were labeled with 4-plex iTRAQ reagents according to the manufacturer's instructions (AB Sciex, Foster City, CA). For each 1 mg peptide from each breast tumor sample, 10 units of labeling reagent were used. Peptides were dissolved in 300 µL of 0.5 M triethylammonium bicarbonate (TEAB) (pH 8.5) solution and labeling reagent was added in 700 µL of ethanol. After 1 h incubation, 1.5 mL of 0.05% TFA was added to stop the reaction. Differentially labeled peptides were mixed and subsequently desalted on 500 mg tC18 SepPak columns. The combined 4 mg iTRAQ samples per experiment were separated into 24 proteome fractions and 12 phosphoproteome fractions using a 4.6mm x 250mm column RP Zorbax 300 A ExtendC18 column (Agilent, 3.5 µm bead size) on an Agilent 1100 Series HPLC instrument by basic reversed-phase chromatography as described previously[113]. Peptides were separated according to their hydrophobicity using solvent A (2% acetonitrile, 5 mM ammonium formate, pH 10) and a nonlinear increasing concentration of solvent B (90% acetonitrile, 5 mM ammonium formate, pH 10). Phosphopeptides were enriched using Ni-NTA superflow agarose beads (Qiagen, #1018611) that were stripped of nickel with 100 mM EDTA and incubated in an aqueous solution of 10 mM FeCl3 (Sigma, 451649) as described previously[117]. For phosphopeptide enrichment a 80% acetonitrile/0.1% trifluoroacetic acid binding buffer and a 500 mM dibasic sodium phosphate, pH 7.0, (Sigma, S9763) elution buffer were used. Enriched samples were desalted on StageTips as described[117] before analysis by LC-MS/MS.

Analysis of tumor samples by high performance liquid chromatography tandem mass spectrometry (LC-MS/MS): All peptides were separated with an online nanoflow Proxeon EASY-nLC 1000 UHPLC system (Thermo Fisher Scientific) and analyzed on a benchtop

Orbitrap Q Exactive mass spectrometer (Thermo Fisher Scientific) equipped with a nanoflow ionization source (James A. Hill Instrument Services, Arlington, MA). The LC system, column, and platinum wire to deliver electrospray source voltage were connected via a stainless-steel cross (360µm, IDEX Health & Science, UH-906x). The column was heated to 50^{o}C using a column heater sleeve (Phoenix-ST) to prevent overpressurizing of columns during UHPLC separation.10% of each global proteome sample in a 2 ul injection volume, or 50% of each phosphoproteome sample in a 4 ul injection volume was injected onto an in-house packed 20cm x 75um diameter C18 silica picofrit capillary column (1.9 µm ReproSil-Pur C18-AQ beads, Dr. Maisch GmbH, r119.aq; Picofrit 10um tip opening, New Objective, PF360-75-10-N-5). Mobile phase flow rate was 200nL/min, comprised of 3% acetonitrile/0.1% formic acid (Solvent A) and 90% acetonitrile /0.1% formic acid (Solvent B), and the 110-minute LC-MS/MS method consisted of a 10-min column-equilibration procedure, a 20-min sample-loading procedure, and the following gradient profile: (min:%B) 0:2; 1:6; 85:30; 94:60; 95;90; 100:90; 101:50; 110:50 (last two steps at 500 nL/min flowrate). Data-dependent acquisition was performed using Xcalibur QExactive v2.1 software in positive ion mode at a spray voltage of 2.00 kV. MS1 Spectra were measured with a resolution of 70,000, an AGC target of 3e6 and a mass range from 300 to 1800 m/z. Up to 12 MS2 spectra per duty cycle were triggered at a resolution of 17,500, an AGC target of 5e4, an isolation window of 2.5 m/z, a maximum ion time of 120 msec, and a normalized collision energy of 28. Peptides that triggered MS2 scans were dynamically excluded from further MS2 scans for 20 sec. Charge state screening was enabled to reject precursor charge states that were unassigned, 1, or >6. Peptide match was enabled for monoisotopic precursor mass assignment.

Protein-peptide identification, phosphosite localization, and quantitation: All MS data were interpreted using the Spectrum Mill software package v5.1 (for comparison with proteomes of human breast tumors from our previous study[83]) and v6.0 pre-release (Agilent Technologies, Santa Clara, CA) co-developed by the authors. Similar MS/MS spectra acquired on the same precursor m/z within +/- 45 sec were merged. MS/MS spectra were excluded from searching if they failed the quality filter by not having a sequence tag length > 0 (i.e., minimum of two masses separated by the in-chain mass of an amino acid) or did not have a precursor MH+ in the range of 750-6000. MS/MS spectra from were searched against a database consisting of RefSeq release 60 containing 31,767 human proteins, 24,821 mouse proteins, and an appended set of 85 common laboratory contaminant proteins (RefSeq.20130727-Human.20130730-MouseNR.mm13.contams). Scoring parameters were ESI-QEXACTIVE-HCD-v2, for whole proteome datasets, and ESI-QEXACTIVE-HCD-v3 parameters were for phosphoproteome datasets. All spectra were allowed +/- 20 ppm mass tolerance for precursor and product ions, 40% minimum matched peak intensity, and trypsin allow P enzyme specificity with up to 4 missed cleavages. Fixed modifications were carbamidomethylation at cysteine. iTRAQ labeling was required at lysine, but peptide N-termini were allowed to be either labeled or unlabeled. Allowed variable modifications for whole proteome datasets were acetylation of protein N-termini, oxidized methionine, deamidation of asparagine, pyro-glutamic acid at peptide N-terminal glutamine, and pyro-carbamidomethylation at peptide N-terminal cysteine with a precursor MH+ shift range of -18 to 64 Da. Allowed variable modifications for phosphoproteome dataset were revised to disallow deamidation and allow phosphorylation of serine, threonine, and tyrosine with a precursor MH+ shift range of 0 to 272 Da.

Identities interpreted for individual spectra were automatically designated as confidently assigned using the Spectrum Mill autovalidation module to use target-decoy based false-discovery rate (FDR) estimates to apply score threshold criteria via two-step strategies. For the whole proteome datasets thresholding was done at the spectral and protein levels. For the phosphoproteome datasets thresholding was done at the spectral level. In step 1, peptide autovalidation was done first and separately for each iTRAQ 4-plex experiment consisting of either 25 LC-MS/MS runs (whole proteome) or 13 LC-MS/MS runs (phosphoproteome) using an auto thresholds strategy with a minimum sequence length of 7(whole proteome) or 8 (phosphoproteome), automatic variable range precursor mass filtering, and score and delta Rank1 – Rank2 score thresholds optimized to yield a spectral level FDR estimate for precursor charges 2 thru 4 of <0.6% for each precursor charge state in each LC-MS/MS run. For precursor charges 5-6, thresholds were optimized to yield a spectral level FDR estimate of <0.3 % across all runs per iTRAQ 4-plex experiment (instead of each run), to achieve reasonable statistics, since many fewer spectra are generated for the higher charge states.

In step 2 for the whole proteome datasets, protein polishing autovalidation was applied separately to each iTRAQ 4-plex experiment to further filter the PSM's using a target protein-level FDR threshold of zero. The primary goal of this step was to eliminate peptides identified with low scoring peptide spectrum matches (PSM's) that represent proteins identified by a single peptide, so-called "one-hit wonders". After assembling protein groups from the autovalidated PSM's, protein polishing determined the maximum protein level score of a protein group that consisted entirely of distinct peptides estimated to be false-positive identifications (PSM's with negative delta forward-reverse scores). PSM's were removed from the set obtained in the initial

peptide-level autovalidation step if they contributed to protein groups that have protein scores below the max false-positive protein score. In the filtered results each identified protein detected in an iTRAQ 4-plex experiment was comprised of multiple peptides unless a single excellent scoring peptide was the sole match. For the whole proteome datasets the above criteria yielded false discovery rates (FDR) of <0.5% at the peptide-spectrum match level and <0.8% at the distinct peptide level for each iTRAQ 4-plex experiment. After assembling proteins with all the PSMs from all the iTRAQ 4-plex experiments together the aggregate FDR estimates were 0.42% at the at the peptide-spectrum match level, 1.5% at the distinct peptide level, and <0.01% (1/11,372) at the protein group level. Since the protein level FDR estimate neither explicitly required a minimum number of distinct peptides per protein nor adjusted for the number of possible tryptic peptides per protein, it may underestimate false positive protein identifications for large proteins observed only on the basis of multiple low scoring PSMs.

In calculating scores at the protein level and reporting the identified proteins, redundancy was addressed in the following manner: the protein score was the sum of the scores of distinct peptides. A distinct peptide was the single highest scoring instance of a peptide detected through an MS/MS spectrum. MS/MS spectra for a particular peptide may have been recorded multiple times, (i.e. as different precursor charge states, in adjacent bRP fractions, modified by deamidation at Asn or oxidation of Met, or different phosphosite localization) but were still counted as a single distinct peptide. When a peptide sequence >8 residues long was contained in multiple protein entries in the sequence database, the proteins were grouped together and the highest scoring one and its accession number were reported. In some cases when the protein sequences were grouped in this manner there were distinct peptides that uniquely represented a

lower scoring member of the group (isoforms, family members, and different species). Each of these instances spawned a subgroup. Multiple subgroups were reported and counted towards the total number of proteins, and were given related protein subgroup numbers (e.g. 3.1 and 3.2: group 3, subgroups 1 and 2). To better dissect the tumor/stroma (human/mouse) origin of orthologous proteins in this xenograft experiment, the inclusion of peptides contributing to each subgroup was restricted by enabling the subgroup-specific (SGS) option in Spectrum Mill. Only subgroup-specific peptide sequences were counted toward each subgroup's count of distinct peptides and protein level TMT quantitation. The SGS option omits peptides that are shared between subgroups. If evidence for BOTH human and mouse peptides from an orthologous protein were observed, then peptides that can not distinguish the two (shared) were ignored. However, the peptides shared between species were retained if there was specific evidence for only one of the species, thus yielding a single subgroup attributed to only the single species consistent with the specific peptides. Furthermore, if all peptides observed for a protein group were shared between species, thus yielding a single subgroup composed of indistinguishable species, then all peptides were retained. Assembly of confidently identified PSM's yielded 20,480 total protein subgroups from 11,372 protein groups. Human and mouse ortholog proteins were typically arranged into individual subgroups.

In step 2 for the phosphoproteome datasets a phosphosite table were assembled with columns for individual iTRAQ 4-plex experiments and rows for individual phosphosites. PSM's were combined into a single row for all non-conflicting observations of a particular phosphosite. (i.e. different missed cleavage forms, different precursor charges, confident and ambiguous localizations, different sample handling modifications). For related peptides neither observations

with a different number of phosphosites nor different confident localizations were allowed to be combined. Selecting the representative peptide from the combined observations was done such that once confident phosphosite localization was established, higher identification scores and longer peptide lengths are preferable. After assembling the phosphosite table a polishing step was applied to further filter the phosphosites with the primary goal of eliminating phosphosites with representative peptides identified through low scoring peptide spectrum matches (PSM's) that were observed in only a few experiments. The initial table of representative peptides for 82,030 phosphosites had an aggregate FDR of 3.3% at phosphosite-level. The table was sorted by identification score and then by number of iTRAQ 4-plex experiments in which the phosphosite was observed. The cumulative FDR trend showed inflection points at an identification score of ~8. Phosphosites with an identification score < 8.0 observed in <3/9 experiments were therefore removed, yielding 68,385 phosphosites with an aggregate FDR of 0.34% at the phosphosite level. While the Spectrum Mill identification score is based on the number of matching peaks, their ion type assignment, and the relative height of unmatched peaks, the phosphosite localization score is the difference in identification score between the top two localizations. The score threshold for confident localization (>1.1) essentially corresponds to at least 1 b or y ion located between two candidate sites that has a peak height 10% of the tallest fragment ion (neutral losses of phosphate from the precursor and related ions as well as immonium and iTRAQ reporter ions are excluded from the relative height calculation). The ion type scores for b-H3PO4, y-H3PO4, b-H2O, and y-H2O ion types are all set to 0.5. This prevents inappropriate confident localization assignment when a spectrum lacks primary b or y ions between two possible sites but contains ions that can be assigned as either phosphate loss ions for

one localization or water loss ions for another localization. In aggregate, 66.3% of the reported phosphosites were fully localized to a particular serine, threonine, or tyrosine residue.

Relative abundances of proteins and phosphosites were determined in Spectrum Mill using iTRAQ reporter ion intensity ratios from each PSM. A protein-level or phosphosite-level iTRAQ ratio was calculated as the median of all PSM level ratios contributing to a protein subgroup or phosphosite remaining after excluding those PSM's lacking an iTRAQ label, having a negative delta forward-reverse score (half of all false-positive identifications), or having a precursor ion purity < 50% (MS/MS has significant precursor isolation contamination from co-eluting peptides). Unless stated otherwise for a particular analysis, the following considerations apply to the tumor/stroma (human/mouse) origin of a protein in this xenograft experiment. For the proteome dataset, only PSM's from subgroup-specific peptide sequences contributed to the protein level quantitation (see protein subgrouping description above). A protein detected with all contributing PSM's shared between human and mouse was considered to be human. For the phosphoproteome dataset, a phosphosite was considered to be mouse if the contributing PSM's were distinctly mouse and human if they were either distinctly human or shared between human and mouse. A 2-component Gaussian mixture model-based normalization approach was used to center the distribution of iTRAQ log-ratios around zero in order to nullify the effect of differential protein loading and/or systematic MS variation[83]. Downstream analyses presented in the main figures were restricted to proteins/phosphosites quantified in at least 10 out of the 24 samples with non-missing values, with the exception of the previously described mRNA-protein correlation analysis[28] requiring quantification in 30%, or 8 out of 24, PDX samples. Specific filtering procedures are noted in descriptions of the relevant methods.

Bioinformatics analyses

Cross data-type integration

All gene names were converted to HUGO Gene Nomenclature Committee's approved gene names for comparison across levels and datasets. For mRNA and protein expressions, expression values were collapsed across transcripts or isoforms to the corresponding gene using the highest mean when there were two transcripts or isoforms, or the value with the highest connectivity when there were more than three transcripts or isoforms as implemented in the WGCNA R package[118].

mRNA-protein correlation

Spearman correlations between normalized RSEM values and protein quantifications were calculated for genes that were observed in at least 30% of samples for both RNA-seq and mass-spectrometry as previously described. KEGG pathway enrichment analysis of the correlation was carried out using a Kolmogorov-Smirnov test, and P values were adjusted using the Benjamini-Hochberg procedure.

Proteomic clustering and subtyping

We first applied filtering for protein and phosphosite markers observed in at least 10 samples and sufficient deviations across samples. We used a 2 standard deviation threshold for iTRAQ proteome and 2.5 standard deviation for the phosphoproteome. We applied the same protein marker to conduct LFQ proteome and PDX vs. human proteome co-clustering. For the co-

clustering with human proteome, we further selected for markers that had higher than 2 standard deviations in the merged proteome and showed non-differential expression between human and PDX (FDR > 0.3, t-test). The subsequent hierarchical clustering was conducted using the complete agglomeration method of hclust as implemented in the heatmap2 R package.

Differential expression analysis

Differential expression testing of each protein in the LFQ and iTRAQ datasets was conducted using the student's t-test, and P values were adjusted using the Benjamini-Hochberg procedure. For gene set enrichment analysis, we conducted the Wilcoxon Rank Sum Test to test for changes in the t-statistics ranks of protein members in each of the KEGG signaling pathway, and again adjusted p values using the Benjamini-Hochberg procedure.

Druggable genes and mutations

We compiled a list of druggable genes based on extensive curation of public databases: Tumor Alterations Relevant for GEnomics-driven Therapy (TARGET, version 3, assessed on 6/15/2015), Personalized Cancer Therapy (PCT, assessed on 3/15/2015), GDKD (Gene-Drug Knowledge Database, version 11.0, assessed on 4/10/2015), CancerDR (assessed on 2/6/2015), My Cancer Genome (assessed on 9/11/2014), and DrugBank (assessed on 9/21/2015). We curated the list based on evidence level and literature, as well as IC50 data when available. The final list used for the analysis included 76 druggable genes.

Druggable outlier analysis

To discover expression outliers, we utilized a strategy incorporating multiple steps: first, we limited our search to genes that are in the druggable gene list. We then narrowed down the list to genes that are observed in at least 10 samples in the dataset under investigation. Outlier expressions were defined as values that are greater than 1.5 interquartile ranges (IQRs) above the third quartile (Q3). To rank order outlier expression for each gene, we calculated an outlier score defined as:

Outlier score = (x –Q3)/IQR

By definition, genes with outlier score greater than 1.5 are considered as expression outliers. Outlier score for each gene were ranked within the sample to select the most promising druggable targets. For CNV outliers, we required them to have outlier scores above 1 in at least another expression level. For validation of the proteomic druggable outliers, we counted the numbers of the same protein or phosphosite outliers observed in the parallel-processed cohort of 77 human breast cancer samples. Due to the lower numbers in this larger cohort, we considered both proteins with outlier score greater than 1 and the top 2 outliers of each human sample as validating outliers.

Pathway activation analysis

We first collapsed the phosphosites to gene-level phosphorylation values by averaging the phosphosite expressions observed for each gene. Then, we converted the phosphoproteomic expression values from iTRAQ to modified z-scores normalized against the cohort as described in Hoaglin et al. (How to Detect and Handle Outliers). We then used the Wilcoxon Rank Sum Test to test for changes in the phosphorylation z-score ranks of protein members in each of the

KEGG signaling pathway. The resulting P values were adjusted using the Benjamini-Hochberg procedure.

Chapter 4: Redefine druggable targets in breast cancer by global phospho-proteomics

4.1 Abstract

Aberrant phospho-signaling is a hallmark of cancer. We investigated kinase-substrate regulation of 33,239 phosphorylation sites (phosphosites) in 77 breast tumors followed by validation in 24 xenografts. Our search discovered 19,521 novel phosphosites and 2,134 total correlated kinase-phosphosite pairs. Among the 91 kinases with auto-phosphorylation, elevated EGFR, ERBB2/MAP3K5, PRKG1, and WNK1 regulations were enriched in basal, HER2-E, Luminal A, and Luminal B breast cancers, respectively, revealing subtype-specific signaling. CDKs, MAPKs, and ataxia-telangiectasia proteins were dominant, master regulators of substrate-phosphorylation whose activities are not captured by genomic evidence. We unveiled activated phospho-signaling, prioritizing targets from 113 activated kinase-substrate pairs and cascades downstream of kinases including AKT1, BRAF and EGFR. We further identified kinase-substrate-pairs associated with clinical biomarkers and phosphoproteomic immune signatures. Overall, kinase-substrate regulation revealed by the largest unbiased global phosphorylation data to date connects driver events to their signaling effects and potentially inform rational targeted treatment of individual breast tumors.

4.2 Results

Catalog of phosphosites in breast cancer

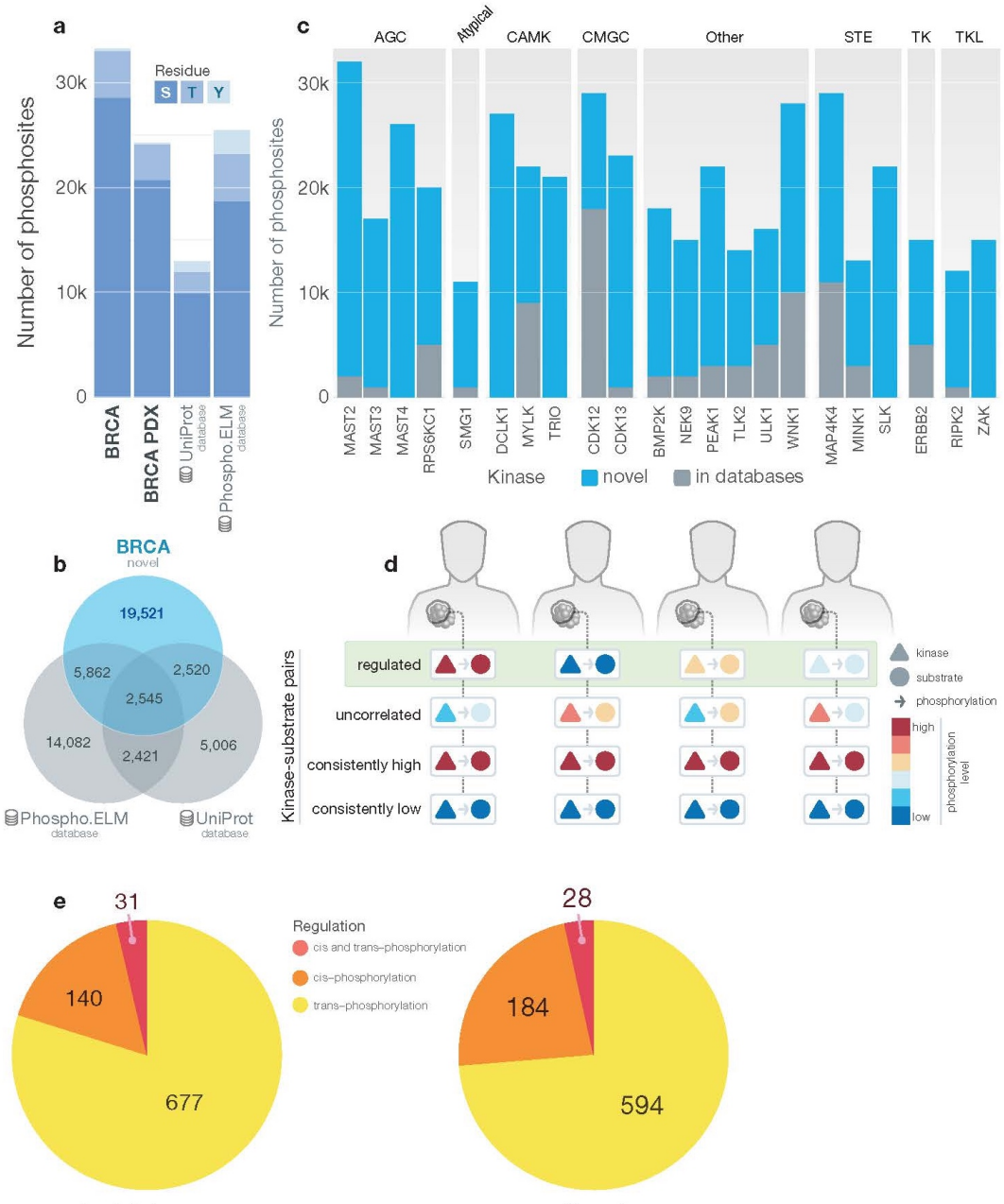


Figure 4.1. Landscape of the 33,239 quantified phosphosites in breast cancer. (a) Distribution of phosphosites and counts of serine (S), threonine (T), and tyrosine (Y) residues

in the 77 breast cancer (BRCA) and 24 PDX cohorts compared to the UniProt and Phospho.ELM databases. (b) Venn diagram showing quantified phosphosites in human tumors compared to phosphosites in the UniProt and Phospho.ELM databases. (c) Number of novel phosphosites and known phosphosties from UniProt and Phospho.ELM detected in kinases of major kinase groups. Only kinases having more than 10 novel sites are shown. (d) Diagram of regulated, uncorrelated, consistently high and consistently low kinase-substrate pairs in breast cancer samples. (e) Pie charts showing the numbers of *cis*-regulated and trans-regulated phosphosites that are novel (left) and previously known (right).

We analyzed genomics and proteomics data from 77 breast cancer and 3 normal breast samples characterized by TCGA[119] and CPTAC[23] (Figure 4.1, Methods). Genomic analysis of TCGA DNA-Seq, SNP-array and RNA-Seq data provided comprehensive assessment of somatic mutations, copy number variations (CNV), and mRNA expression, respectively. The CPTAC breast cancer project quantified global protein and phosphosite expression levels using the Isobaric Tag for Relative and Absolute Quantitation (iTRAQ) technology. We further utilized proteogenomic data from 24 independent breast cancer patient-derived xenografts (PDXs) [120] generated using the same technologies and pipeline for validation.

As previously described[83], 33,239 out of the 62,679 confidently identified phosphosites passed additional missing value and standard deviation filter and were used for downstream analyses (Methods). This catalog of breast cancer phosphosites added 19,521 novel phosphosites when compared to the 12,952 and 24,911 genomically-mapped human phosphosites of the UniProt and the Phospho.ELM database[121] (Figure 4.1a,b). These phosphosites covered 315 out of the 523 previously described human kinases[122,123]. Kinases with a significant number of novel phosphosites may have an underappreciated regulatory role in breast cancer, including MAST2

(30 novel phosphosites) and MAST4 (26) from the AGC kinase group, CDK13 (22) and CDK12 (11) from the CMGC kinase group, SLK (22) and MAP4K4 (18) from the STE kinase group, and ERBB2 (10) from the TK group (Figure 4.1c). Compared to other quantified sites, the 84 detected, known cancer-associated phosphosites (Methods), including AKT1 p.T308, GSK3B p.S9 and MTOR p.S2448, exhibited higher standard deviation in both breast cancer and PDX samples (Wilcoxon rank sum test, $p = 0.002567$ and 0.0002608, respectively).

In this study, we aim to identify correlated kinase-substrate pairs that are concordant in their relative abundance across samples (Figure 4.1d). We hypothesize these phosphosites are regulated in a patient-specific manner, contributing to cross-sample variation in wiring of signaling networks. We examined two classes of kinase-substrate relations: (1) *cis* interactions whereby a kinase protein expression is correlated with its own phosphosites and (2) *trans* interactions whereby a kinase phosphoprotein expression is correlated with a substrate phosphosite level (Methods). To achieve this we curated and screened 4,997 pairs of human kinase and substrate proteins based on the PhosphositePlus and PhosphoNetwork databases[124,125]. 806 novel and 848 previously identified phosphosites are identified as regulated: 324 by *cis*-regulation, 1,271 by *trans*-regulation and 59 by both interactions (Figure 4.1e).

Phosphosites in auto-phosphorylated kinases

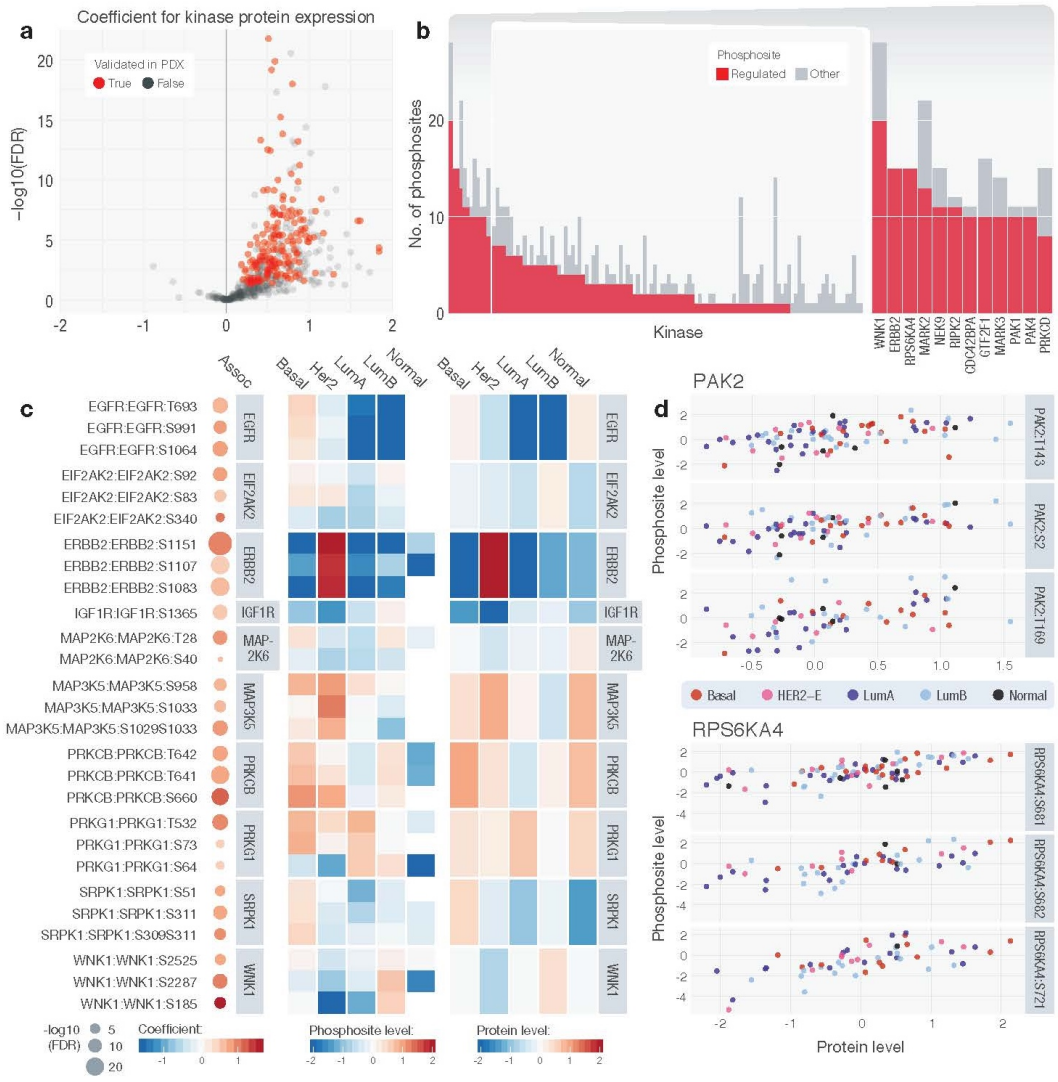


Figure 4.2. Regulated *cis* kinase-phosphosite pairs. (a) Volcano plots highlighting positively cis-regulated pairs. Regulations that were validated in the 24 PDX models ($P < 0.05$) are colored in red. (b) Counts of *cis*-regulated phosphosites among all corresponding phosphosites for each kinase. Kinases with more than 7 *cis*-regulated phosphosites are shown individually in a close-up barplot (right). (c) Top associated *cis* pairs and their average kinase protein and phosphosite levels in each of the breast cancer subtypes. On the left panel, each

dot represents a *cis* association identified by the regression analysis, where its size repr
the significance and color represents the regression coefficient. The middle and right p
illustrate the average *cis*-regulated phosphosite expression and kinase protein expressio
respectively, in each of the breast cancer subtypes. (d) Correlations of kinase protein le
PAK2 and PRS6KA4 and their respective top 3 *cis*-regulated phosphosites. Each dot
represents one tumor sample colored according to its subtype.

For *cis* analysis, protein expression of 120 kinases known to exhibit auto-phosphorylation were evaluated with relation to peptide abundance of their 630 phosphosites using a linear regression model (Methods). Protein abundance measures do not always guarantee the activation of the phosphosites of the same protein: 61.4% (387/630) of the tested kinase-substrate relations showed significant positive associations (regression coefficient $\beta > 0$ and FDR < 0.05) in breast cancer (Figure 4.2a). Out of the 387 *cis*-regulated sites, 348 (89.9%) are novel kinase target sites absent in PhosphositePlus. These novel *cis*-regulated sites are in 98 unique kinases including ERBB2, EGFR and MAP3K5. Of the identified *cis*-regulatory pairs, 98.4% with sufficient PDX data (179/182) showed positive correlation ($\beta > 0.1$) in the cohort of 24 breast cancer PDXs[120] and were validated. The AGC kinase group was most significantly enriched with *cis* associations (Fisher's Exact Test, $P = 0.0241$, Methods) with all 17 tested genes showing significant cis correlations with its phosphosites, followed by the STE group (14/14 genes, $P = 0.0487$). Notably, at the kinase family level, we observed significant *cis*-regulation in all 8 of the STE20 kinases included in the analysis including PAK1/2/4, STK3/4/24 and MAP4K1/3.

Top proteins with high percentage of *cis*-regulated phosphosites (Figure 4.2b) included well-known and novel breast cancer associated-proteins such as ERBB2 (15 significant *cis*-regulated

phosphosites), RPS6KA4 (15), NEK9 (11), RIPK2 (11), CDC42BPA (10) and GTF2F1 (10). Further, at least 7 *cis*-regulated phosphosites in each of ERBB2, PAK4, NEK9, and RIPK2 were validated in the PDX cohort ($\beta > 0.1$). Various *cis*-regulations exhibited molecular distinctions across breast cancer subtypes (Figure 4.2c), exemplified by elevated ERBB2 protein and phosphorylation in HER2-E breast cancers. Other subtype-specific regulated pairs identified were WNK1 up-regulated in luminal A, and EGFR and SRPK1 up-regulated in basal breast cancers. Closer examination revealed clear correlations between levels of PAK2 protein and its p.T143 and p.S2 phosphosites as well as PRS6KA4 protein and its p.S681 and p.S682 phosphosites (Figure 4.2d).

We hypothesized that the observed *cis* associations were mainly due to two mechanisms: (1) A higher protein level directly increases the abundance phosphorylation site level, where each unit change in kinase expression would result in an equally proportional change in phosphosite level, and (2) The kinase autophosphorylates itself, where each unit change in kinase protein may result in a higher unit change in phosphosite level. We identified 41 *cis*-regulated sites that may be affected by auto-phosphorylation using this criterion ($\beta > 1$), including CDK9 p.T186, RAF1 p.S43, BRAF p.S151/S729 and PRS6KA4 p.S737.

Kinase-substrate interactions

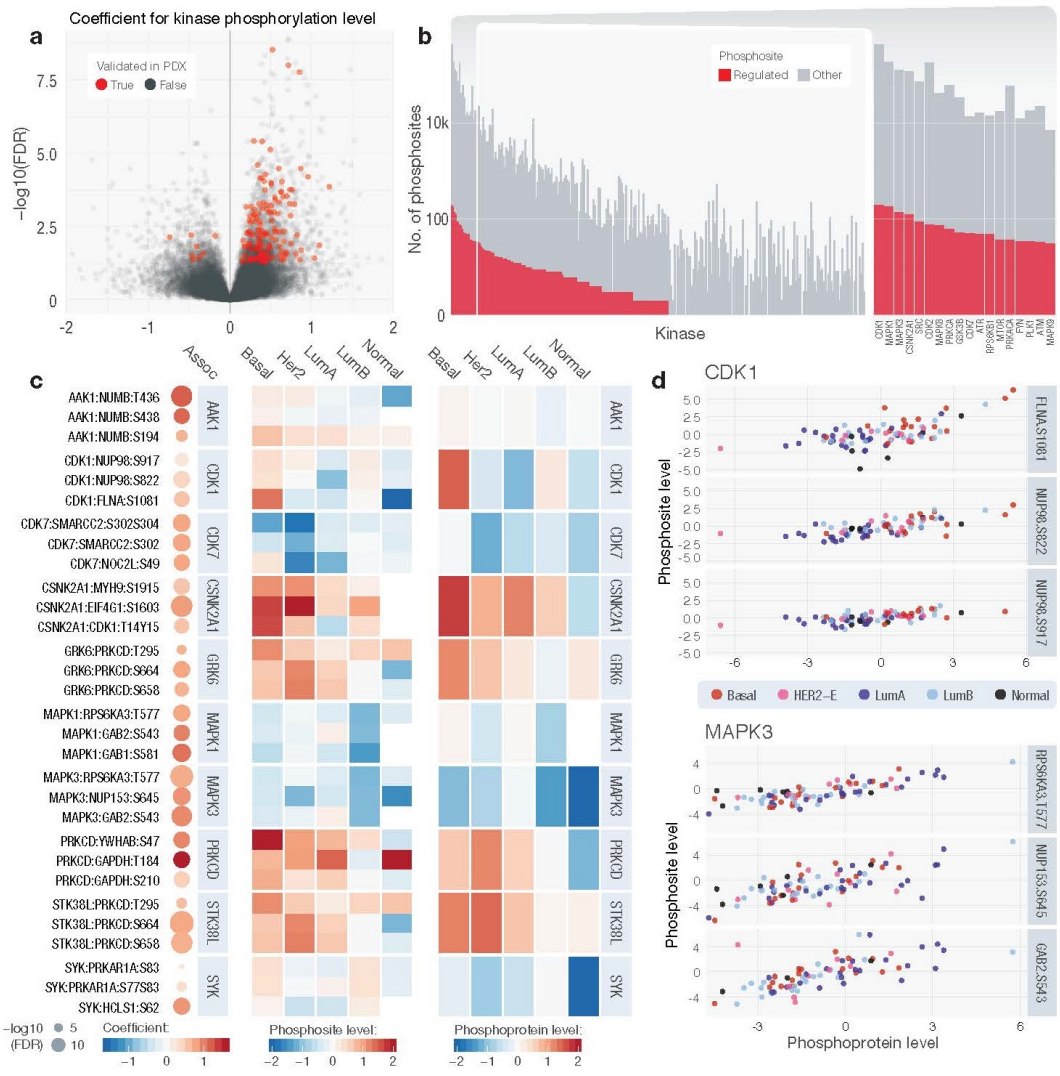


Figure 4.3. Regulated *trans* kinase-substrate phosphosite pairs. (a) Volcano plots highlighting positively regulated pairs. Regulations that were validated in PDX models ($P < 0.05$) are colored in red. (b) Counts of *trans*-regulated phosphosites among all corresponding phosphosites for each kinase. Kinases with more than 30 *trans*-regulated phosphosites are shown individually in a close-up barplot (right). (c) Top associated *trans* pairs and their average kinase phosphoprotein and substrate phosphosite levels in each of the breast cancer subtypes. On the left panel, each dot represents a *trans* association identified by the

regression analysis, where its size represents the significance and color represents the regression coefficient. The middle and right panels illustrate the average *trans*-regulate substrate phosphosite expression and kinase protein expression, respectively, in each o breast cancer subtypes. (d) Correlations of kinase phosphoprotein level of CDK1 and MAPK3 and their respective top 3 *trans*-regulated phosphosites. Each dot represents o tumor sample colored according to its subtype.

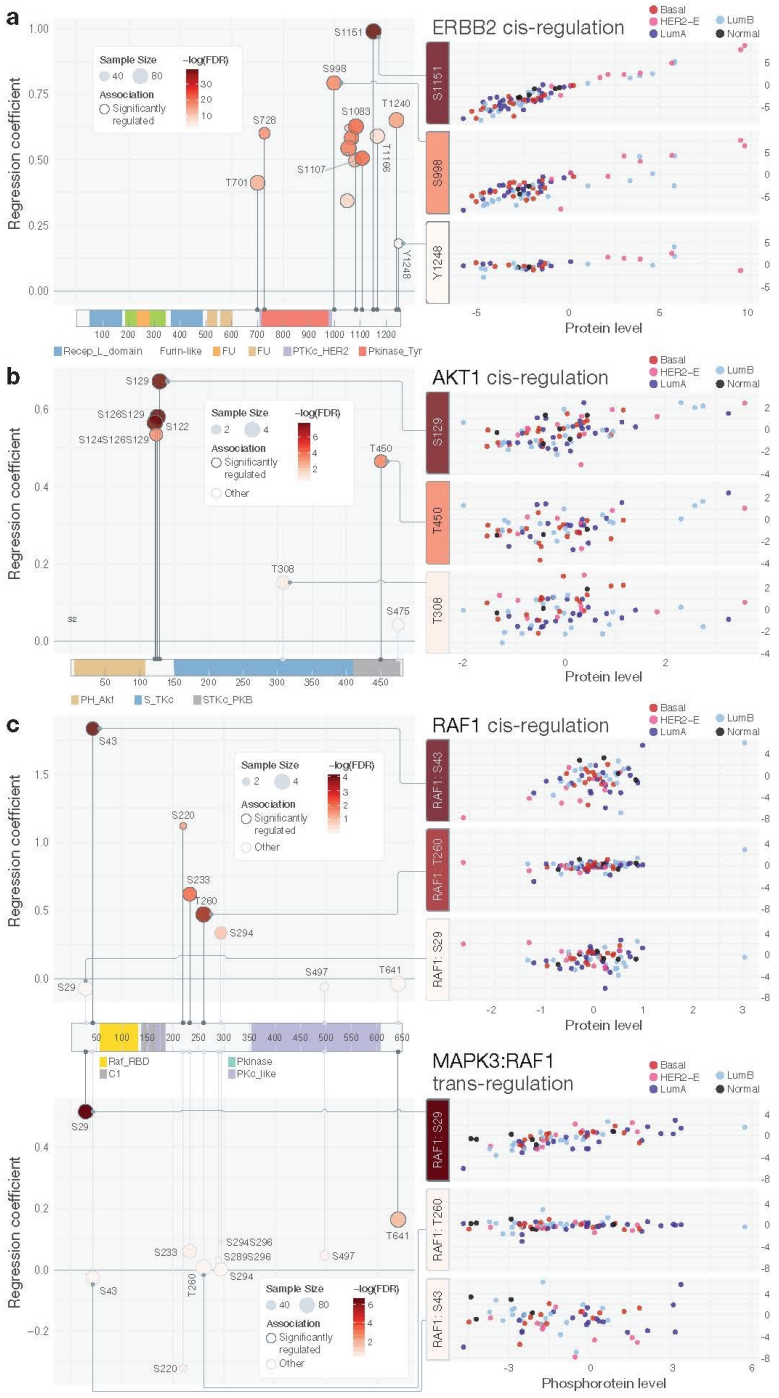
For *trans* analysis, we surveyed 7,404 kinase-substrate protein pairs and a total of 38,710 kinase-substrate phosphosite relations with sufficient observations in our data set (Methods). We applied a linear model using phosphosite abundance of the substrate as the dependent variable and phosphoprotein expression of the kinase as the independent variable and the protein expression of the substrate as a covariate. Only 4.51% (1,747/38,710) of the tested relations showed significant positive associations (Figure 4.3a). Among the 1287 *trans*-regulatory pairs based on kinase-substrate relations of PhosphositePlus, 1117 (86.8%) are novel sites absent in the existing databases. A significant fraction (45.8%) of these novel sites are regulated by CDKs, MAPKs, ATR, and ATM. Of the identified *trans*-regulatory pairs, 45.5% with sufficient PDX data (407/894) were confirmed ($\beta > 0.1$) in breast cancer PDXs[120]. The moderate validation rate may be due to increased sensitivity to micro-environment in *trans* pairs and larger sample sizes may be required to firmly establish *trans* associations. The identified *trans*-regulations are congregated in 165 kinases, which showed the most significant enrichment in MAPK ($P = 0.000448$), CDK ($P = 0.00259$), and PKC ($P = 0.00613$) families.

Multiple kinases with the highest number of *trans* associations have been previously implicated in breast cancer. Particularly, ATM (33 *trans*-regulated substrate phosphosites) and ATR (44), proteins known to orchestrate the DNA damage response (DDR) pathway, were significantly

associated with phosphosites of chromatin-associated SMARCC2, MCM2, histone-lysine methyltransferase KMT2A and the DNA damage checkpoint protein TP53BP1. In the PI3K pathway, kinases associated with *trans*-regulations of more than 20 phosphosites included GSK3B (35), RPS6KB1 (39), MTOR (29) and RPS6KA1 (21). We also observed specific interactions such as the inhibitory phosphorylation of GSK3A by AKT1 (AKT1:GSK3A p.S278, FDR = 0.0341)[126].

CDK1 (178 *trans*-regulated phosphosites), CDK2 (72) and CDK7 (44) showed a wide-spread effect of *trans*-regulations on substrates including NUP98 and FLNA, supporting their central roles in cell cycle signaling. MAPK1 (148 *trans*-regulated phosphosites), MAPK3 (111), MAPK8 (50), MAPK14 (18) and MAPK9 (15) were also associated with up-regulation of phosphosites of multiple downstream substrates including RPS6KA3/5, GAB1/2, MAP2K1, and CIC (Figure 4.3c,d).

Sequence and structural patterns of phosphosites



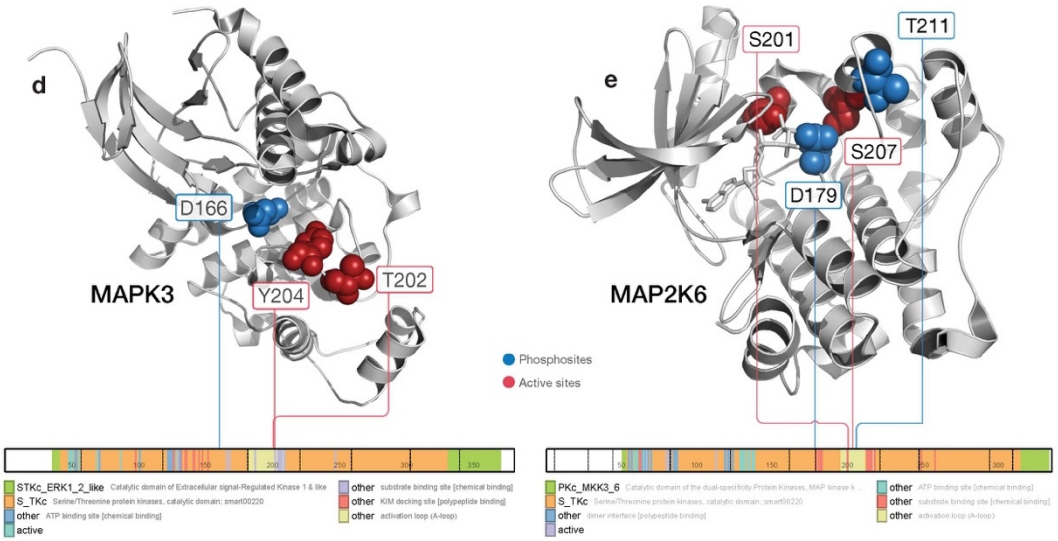


Figure 4.4. Patterns of regulated phosphosites on primary sequences and 3D struc

(a) Consistent *cis*-regulations of phosphosites idenfied in ERBB2. (b) Discordant *cis*-regulations of phosphosites identified in AKT1. (c) *Cis* and *trans*-regulations by MAPI RAF1 phosphosites. (d) *Cis*-regulated phosphosites p.T202 and p.Y204 in spatial prox adjacent to the active site p.D166 of MAPK3 as in PDB structure 4QTB [127]. (e) *Trans*-regulated phosphosites p.S201 and p.S207 (by MAP3K5) are found in spatial proximit the active sites, including p.D179 and p.T211, of MAP2K6, which is co-crystalized wi ATP analog as in PDB structure 3VN9 [128].

Our *cis* and *trans* analyses identified pairs of phosphosites on the same protein showing concordant regulation patterns, including those in ERBB2, PRKCB, and WNK1. We hypothesized that phosphosites in spatial proximity would be affected by the same regulators and show similar phosphorylation patterns. For each pair of phosphosites, we calculated the correlation coefficient of their levels and compared that to their distances on PDB structures as determined by HotSpot3D[74,126]

(Methods). We found a significant association between the correlation of phosphosite levels and 3D distances of the phosphosite pairs (Spearman correlation, P = 3.514e-08), and linear distances (P = 0.000793), confirming co-regulation in adjacent phosphosites.

Phosphosites with strong *cis*-regulations may help detect activating events correlated with high kinase expression. We identified ERBB2 p.S1151/S998/T1240 showing the most significant *cis*-regulations ($\beta > 0.65$; FDR < 6.24E-16) and may serve as complementary HER2 biomarkers to the tyrosine residues targeted by available antibodies, such as p.Y1221 and p.Y1248 (Figure 4.4a). On the other hand, poorly correlated phosphosites in kinases with strong *cis*-effects may suggest additional post-translational modification mechanisms. For example, we observed strong association between AKT1 protein and phosphorylation levels for sites p.S122/S126/S129 (FDR < 8.22E-08). However, the associations for p.T308/S475 (FDR > 0.378) are considerably weaker even with similar observed sample sizes (Figure 4.4b). Other discordant sites include the strongly associated ABL1 p.S637/S737/T800/T871 vs. the non-associated p.S16/S588/S828; the strongly associated PTK2 p.S390/S570/S708/S910 vs. the non-associated p.Y576/S840; and the associated RIPK1 p.S320/S610 vs. the non-associated p.S330/S416.

In RAF1, we observed strong *cis*-regulations at p.S43/T260/S233/S220 (FDR < 0.0358), but not p.S29 (FDR = 0.835, Figure 4.4c). This association pattern is complemented by the *trans*-regulation of RAF1 by MAPK3: RAF1 p.S29/T241 are tightly correlated with MAPK3 phosphoprotein expression (FDR < 0.0320) when controlled for RAF1 protein expression, indicating MAPK3 may specifically regulate RAF1 phosphorylation at these sites via a feedback loop.

Intriguingly, several regulated phosphosites reside in structural proximity to active sites of kinases. MAPK3 (ERK1) p.T202/Y204, which are known to be phosphorylated by MAP2K1 and MAP2K2 (MEK1 and MEK2) to trigger its activation[129,130], are adjacent to its active site p.D166 (Figure 4.4d). These two sites showed significant *cis*-regulation by MAPK3 protein albeit not *trans*-regulation by MAP2K1/2 phosphoprotein. In MAP2K6, MAP3K5-regulated p.S207 and an adjacent site p.S201 are located in its catalytic domain near active sites p.D179 and p.T211 (Figure 4.4e). These phosphosites may alter the biochemical properties of the active site and affect the activity level of the corresponding kinase.

Kinase-substrate pairs refine treatment options

Figure 4.5. Druggability analysis of single and paired events in 77 breast cancer samples. (a) Heatmap of regulated kinase-substrate pairs where the kinase or the substrate is a potential druggable target. The sample-pair showing outlier pair event is outlined. Only the

top 3 regulated pairs were shown for each kinase when there were more than 3 pairs sh

To inform targeted treatment, kinase-substrate outliers. (b) Druggable events identified in the mutation, CNV, RNA, protein, and phospho-pair level for breast cancer samples in the same subtype order. druggability analysis has traditionally focused on detecting single activating events, such as EGFR mutation, PIK3CA mutation, and ERBB2 amplification[13,92]. While these events alone can predict treatment response, co-occurrence of elevated kinase-substrate phosphorylation can confirm aberrant activation and reveal new treatment options. We conducted druggability analysis for the significant *cis* and *trans* kinase-substrate pairs by screening 68 druggable genes with potential inhibitors [120] that are abundantly expressed in our breast cancer samples (Methods). We used abundance levels of both the kinase and substrate phosphosite to construct a kinase-substrate pair score and identified samples showing 2 standard deviations above cohort medians in their scores (Methods).

Among the 286 associated kinase-substrate pairs, we identified 164 outlier events among 113 unique pairs (Figure 4.5a). Outliers of the *cis* pair ERBB2:ERBB2 p.S1151 were found in 5 HER2-positive samples. The *cis* BRAF:BRAF p.S364 pair and the trans MAPK14:FOXO3 p.S413 pairs were found exclusively in 4 and 3 luminal B breast cancers, respectively. SRC and SYC-regulated *trans* pairs also showed higher levels in luminal B breast cancers. Pairs associated with MAP kinases, such as MAPK3:PALLD p.S55, MAPK3:IL16 p.S584 and MAPK8:JUN p.T62, showed higher levels and several pair outliers in luminal A breast cancers. In samples without prominent ERBB2 signaling, we observed other alternative outlier pairs triggered by other receptor tyrosine kinases, including EGFR regulated *trans* pairs and IGF1R-regulated *cis* pair IGF1R: IGF1R p.S1365. In the 24 breast cancer PDXs, we also discovered

outliers of many regulated pairs including ERBB2:ERBB2 p.S1151, BRAF:BRAF p.S364 and EGFR:PLCG1 p.S1222, supporting their prevalence in breast cancer.

We compared the candidate targets identified through paired druggability analysis and conventional single driver analysis (Figure 4.5b), where we compiled mutations, CNV, RNA, protein levels of the same 68 expressed, potentially druggable genes (Methods). We observed concordant single driver events with outlier pair events for kinases including AKT1, BRAF, ERBB2, IGF1R and RAF1. Samples with *ERBB2*, *IGF1R* and *RAF1* copy number amplified outliers often show high expression of the respective kinases-substrate pairs (Figure 4.5b). However, single driver events do not guarantee activated signaling of the kinase-substrate pairs for these kinases and phosphoproteomics data confirms abberant regulation. For example, some samples with *PIK3CA* mutation, *ERBB2* mutation, or *RAF1* copy number amplification did not show activated phospho-signaling (Figure 4.5b). To further confirm the effect of these effects and treatment option for each patient, observing concurrent activation of downstream targets may be required.

On the other hand, active signaling events may occur in samples without mutations or expression aberrations of the kinase. This is particularly evident in both human and PDX samples with outlier MAPK3 and MAPK14 *trans* pairs (Figure 4.5b). Only 2 out of 7 samples with outlier MAPK3 *trans* pairs showed *MAP3K1* mutations and all 5 samples with outlier MAPK14 *trans* pairs did not carry *MAP3K1* or *MAP2K4* mutations, suggesting MAP kinases may be activated by other upstream signaling mechanisms rather than directly being altered at the sequence or

expression levels. We also observed some of these outlier pair events in absence of single driver events for AKT1, BRAF and EGFR (Figure 4.5b) that require further investigation.

Activated kinase-substrate cascades

We further extended the analysis from pairs to two-level signaling cascades that include the first and second degree substrates of the druggable kinase (Methods). We hypothesize in these signaling cascades, the activated kinases could trigger phosphorylation of various targets through multiple steps and represent good opportunities for targeted inhibition treatment.

Out of 28 kinases associated with at least one *cis* or *trans*-regulated phosphosite, 16 also had second-degree substrates. AKT1, BRAF, EGFR, JAK2, PRKCE and PLK1 showed *cis* and *trans* interactions that may help confirm activity levels (Figure 4.6). In addition to its *cis* associations, AKT1 also associated with phosphosites of ILF3, ETS1 and GSK3A, and GSK3A is in turn associated with 3 phosphosites of NDRG1, 2 sites of RICTOR and 1 site each in LARP1 and ANKS1A (Figure 4.6a). BRAF also has multiple *cis*-regulated phosphosites and is associated with phospho-MAP2K1/2 (MEK1/2), which are associated with downstream phosphosites at MTA1, KRT8 and PRKCD (Figure 4.6b). Finally, high levels of EGFR phosphorylation were mostly observed in basal subtype breast cancers, and these tumors also exhibited high phosphosite levels in GAB1, PLCG1, FAM129B and PTK2; PTK2 phosphoprotein is further associated with phosphosites of PPP1R13L and PXN (Figure 4.6c). The association of the primary druggable kinase with its signaling cascade could strengthen the rationale for targeted inhibition in tumors showing co-activation.

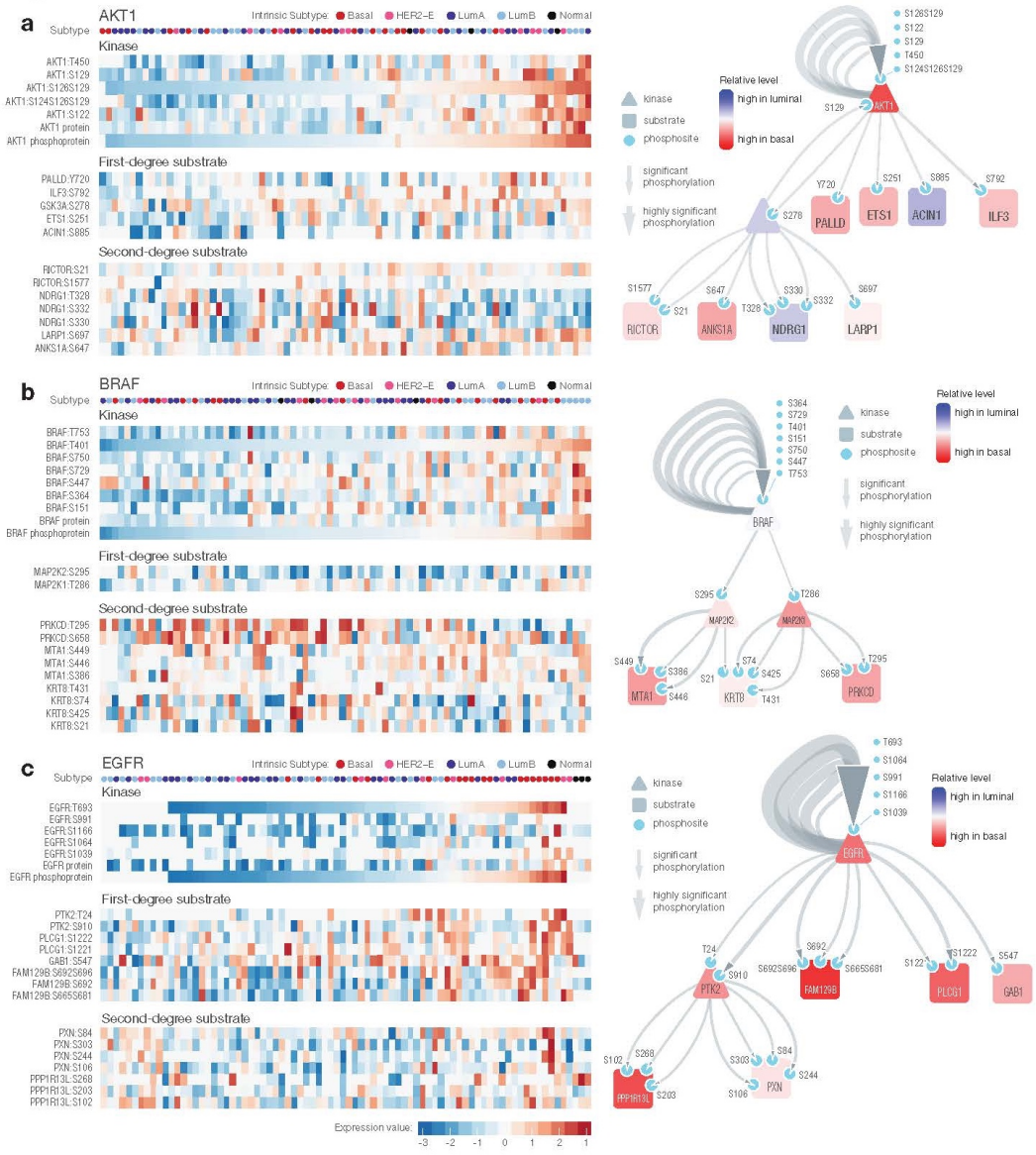


Figure 4.6. Druggable kinase-substrate cascades originating from (a) AKT1 (b) BRAF and (c) EGFR in the 77 breast cancer samples. The samples in the heatmap were ordered by the phosphoprotein level of each of the kinases. For each node in each network diagram, the color represents the relative level of basal compared to luminal A/B breast cancers, where blue indicates higher level in luminal and red indicates higher level in basal tumors. For the

edges, the darkness of the color is scaled by the correlation coefficient and the width is by -log(FDR) of the association.

Kinase-substrate pairs correlated with clinical and immune features

Lastly, we sought to identify kinase-substrate pairs associated with clinical characteristics of breast cancer. We conducted regression analysis between the correlation score for each kinase-substrate pair and pathogical stage and survival adjusted for their PAM50 subtypes (Methods). 175 pairs showed potential association ($P < 0.05$) with pathogical stage (Figure 4.7a); pairs stemmed from MAP kinases, including MAPK11:PPP1R13L p.S113, MAPK9:STAT3 p.S727 and MAPK8:NFATC1 p.S359, showed top correlations with earlier clinical stage. 83 pairs are potentially associated with survival using the Cox proportional hazards model (Figure 4.7b). Notably, these pairs are dominated by 32 CDK1 and 29 CDK2 *trans*-regulated pairs correlated with protective effects (Hazard Rate Ratio < 1) in breast cancer survival.

We identified 160 pairs potentially associated ($P < 0.05$, 40 pairs with FDR < 0.05) with transcriptomically-derived immune score as calculated by the ESTIMATE algorithm[131] (Methods, Figure 4.7c,d). We then specifically asked whether the associated pairs showed direct overlap with the immune-related genes. Our analysis showed that 15 pairs positively associated with the immune scores have their respective substrates as immune genes; the significant enrichment of immune genes in identified pairs (15/160 vs. 11/918, Fisher's Exact Test, P = 3.58e-07) validated our approach. The remaining 145 pairs (the top 6 pairs shown in Figure 4.7d) represent signaling networks simultaneously activated with up-regulation of the immune genes and functions.

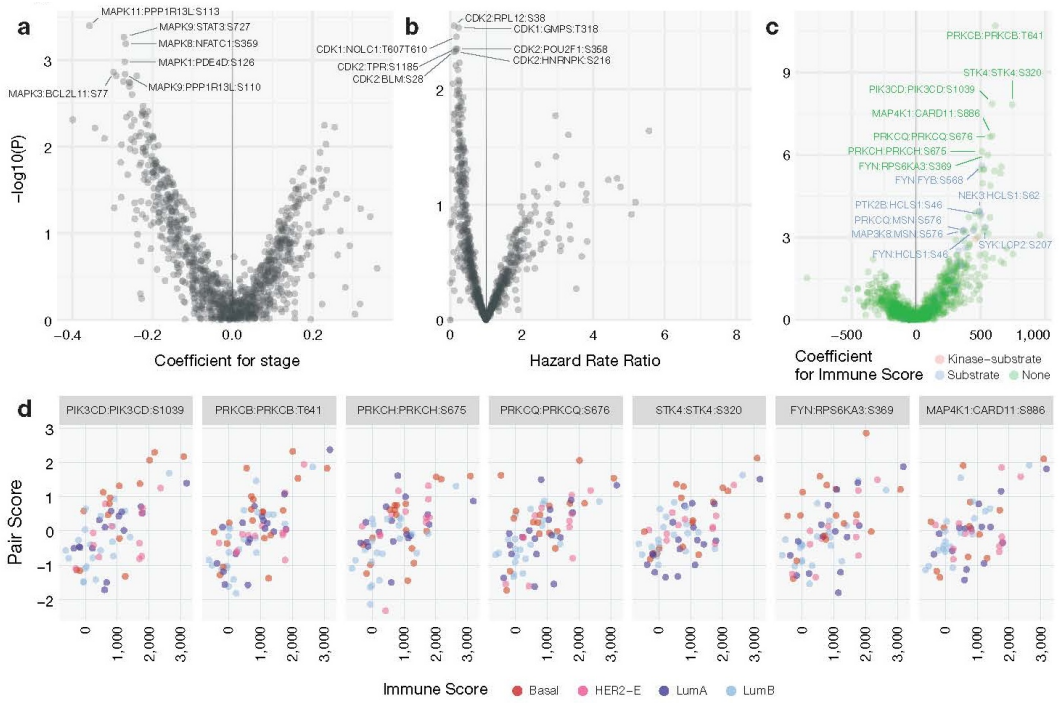


Figure 4.7. Clinical association of kinase-substrate pairs. (a) Volcano plot showing association of kinase-substrate pairs with pathological stage. Positive coefficient denotes higher kinase-substrate scores associating with more advanced pathological stage. (b) Volcano plot showing association of kinase-substrate pairs with survival. Hazard rate ratios greater than 1 denote higher kinase-substrate scores associating with poor survival. (c) Volcano plot showing association of kinase-substrate pairs with transcriptome-based immune signature score, as calculated by the ESTIMATE algorithm. Positive coefficients denote higher kinase-substrate scores associating with higher immune scores. The color of each pair indicates whether its kinase or substrate belongs to the immune gene list used by ESTIMATE. (d) Top kinase-substrate pairs (P < 1e-6) associated with immune scores where

each dot indicates one breast cancer sample.

4.3 Discussion

Herein, we present a quantitative characterization of kinase-substrate pairs in breast cancer (Figure 4.1d). The high-throughput dataset generated by LC-MS/MS enabled global assessment of 33,239 phosphosites, 19,521 of which were not observed in two of the most comprehensive phosphosite databases, UniProt and Phospho.ELM (Figure 4.1a,b). Our analysis allowed us to identify 2,134 (387 *cis* and 1,747 *trans*) kinase/substrate regulatory relationships; using the same pipeline, analysis based on RPPA-detected phosphosites only found 4 and seldom interrogated more than one phosphosite on a protein, further stressing the need of evaluating interactions *in vivo* through global phosphoproteomics. Strikingly, our study discovered 806 novel regulated phosphosites (Figure 4.1e) from a wide spectrum of genes and gene families, regulated by various protein kinases. This result clearly suggests that more regulated, likely cancer-specific phosphosites will emerge in additional, even larger mass spectrometry based cancer proteome studies. While our analyses advances towards a more comprehensive cataloging of phospho-regulations, expanding the current sample size would be required to fully establish and discover associations (Figure 4.7). The serine-rich dataset in this study may also be complemented by other techniques enriching for tyrosine residues[132-134].

We identified 61.4% (387/630) phosphosites of kinases showing significant *cis*-regulation, many of which were concentrated in known or nominated breast cancer genes such as ERBB2, RRPS6KA4, NEK9, RIPK2 and PAK1 (Figure 4.2). In contrast, only 4.51% (1,747/38,710)

trans kinase-substrate pairs showed significant co-regulation (Figure 4.3). It is possible that *trans* substrate usage may be highly tissue-specific and some of the previously curated pairs do not interact in breast cancer. Another possibility is that some kinase-substrate pairs established through *in vitro* evidence are not relevant in physiological environments, although the validation rate for *in vivo* and *in vitro* pairs do not differ significantly. Future investigation using data across tissue and cancer types would be pivotal in addressing whether we observe tissue-specific usage of kinase-substrate pairs. Such studies could also reveal the consistently high/low pairs (Figure 4.1d) in each cancer type and highlight cancer-specific signaling.

Our direct, quantitative observation of kinase-substrate pairs complement previous studies focusing on singleton drivers. Conventionally, pathways were mostly constructed by linking single candidate driver genes (such as significantly mutated genes or focally amplified genes) through known interactions. Our approach detects the co-regulation of the gene pairs *in vivo*, and thus directly validates the signaling impact of driver events in each sample. This approach also enabled us to build relevant sub-cascades stemming from potentially druggable kinases AKT1, BRAF and EGFR (Figure 4.6). To compare with other network-generating studies, we also constructed a network of all observed regulations. However, such approaches may obfuscate activated subnetworks as downstream phosphorylation targets could be mediated by multiple kinases.

This first large-scale examination of over 33,000 phosphosites in breast cancers sets a foundation for druggable analysis of kinase-substrate pairs beyond singleton druggable events (Figure 4.5). Predictive value of response to targeted treament has been limited in samples for some clearly-

definied driver events in cancer (ex. PIK3CA mutation status failing to predict treatment response to PI3K inhibitor). Co-occurrence of downstream activating events, as we have observed for AKT1, BRAF, ERBB2, IGF1R and RAF1 (Figure 4.5b), may further support targeted inhibition. In both breast cancer PDXs showing ERBB2:ERBB2 p.S1151 *cis* outlier pairs, lapatinib treatment significantly reduced tumor growth[120]. Resistance mechanisms often consists of rewiring of signaling pathways and could be further explored through high-throughput proteomics and approaches developed in this study. Further, we identified outlier kinase-substrate pairs in samples without singleton events for kinases including EGFR, MAPK3 and MAPK14 (Figure 4.5b). Inhibition of the MAP2K1/2 (MEK1/2) upstream of MAP kinases surpressed the MAPK signaling pathway and its combinatory treatment with RTK inhibitors have resulted in tumor regression of triple-negative breast tumors[29]. Our discovery of MAPK mediated pairs reveals therapeutic opportunities.

In conclusion, signaling networks are crucially important in cancer. However, large-scale omic studies to date have mainly focused on singling out individual driver events and rarely investigate their signaling impact. Studying kinase-substrate relations *in vivo*, and most particularly in tumor samples from patients undergoing therapy will uncover the wiring of signaling networks in each tumor and likely improve treatment approaches.

4.4 Methods

Sample Description

Samples of human breast cancer were as described in the CPTAC marker paper[23,27]. These comprise of 77 breast cancer samples that showing unimodal distribution in proteomes, their 3 technical replicates and 3 normal breast samples. Samples of the 24 PDX breast cancer were as described previously[120].

Data Generation

TCGA genomics data

The TCGA somatic mutation data, level-3 segment-based copy number data, level-3 normalized RNA expression data, were downloaded from firehose (archive date 2014-10-17). We then converted the segment-based copy number data to the gene-based copy number data by using the RefSeq database (version 20130727). The CNV ploidity number is divided by 2 and then log2-transformed to obtain the final CNV levels for analysis. We also log2-transformed the RSEM values of RNA expression data.

TCGA RPPA data

Normalized RPPA data of TCGA tumors were downloaded from The Cancer Protein Atlas (TCPA , archive date 2015-10-30). The RPPA data were normalized across batches using replicates-based normalization (RBN) as previously described[135].

Global Proteomics data

Global proteomics data for the human samples were downloaded from the Mertins et al. breast cancer study[83]. Global proteomics data for the PDX samples were downloaded from the Huang et al. breast cancer PDX study[120]. As previously described, 2-component Gaussian mixture model-based normalization algorithm was used to normalize the data and accomandate both consistently and differentially-expressed proteins and phosphosites within each sample. Further, proteins and phosphosites were required to have observed (non-missing) iTRAQ ratios in at least 30 samples and an overall standard deviation larger than 0.5 (across all samples where they were observed).

Protein and phosphorylation databases

UniProt: We applied HotSpot3D (v1.1.1) which accesses crystal structures from RCSB Protein Data Bank (PDB) and calculates residue distances using the average distance-measure option in preprocessing (structures processed January 2017)[74]. We used a custom Perl script to retrieve phosphosites, active sites, and binding sites (ie. features matching MOD_RES, ACT_SITE, BINDING, or SITE) from UniProt (date accessed 1/30/2017)[136].

PhosphositePlus: We downloaded the PhosphositePlus[124,137] database (2/11/2016). We then extracted the kinase-substrate pairs where both are proteins are from human from the kinase-substrate database for downstream analysis.

To identify phosphosites known in cancer, we filtered the disease-associated sites database for cancer terms (ex. cancer, *oma and leukemia). The sites that were not matched to a valid genomic coordinate by transvar were excluded, and the remainder sites were further reviewed to retain 261 sites, where 84 unique sites were quantified in our dataset.

Phospho.ELM: We downloaded the phospho.ELM database from PhosphositePlus (2/11/2016). We then extracted the phosphosites mapping to human proteins, and reverse-translated to their unique genomic positions using transvar.

PhosphoNetwork: We downloaded the Supplementary Tables from Newman et al.[31] and derived the predicted kinase-substrate pairs from the file comKSI.csv. We then further filtered out the pairs already observed in PhosphositePlus and combine the remaining pairs with pairs from PhosphositePlus for analysis.

Bioinformatics analyses

Cross data type and database integration

All gene names were converted to HUGO Gene Nomenclature Committee's approved gene names for comparison across levels and datasets. To match the exact phosphosite (ex. PIK3CA:NP_006209.2:s312) across databases, all phosphosites are reversely-mapped to their genomic position (ex. chr3:g.178921452_178921454) using transvar [71].

Regulated kinase-substrate pairs regression analysis

We obtained 3,245 unique kinase-substrate pairs in the PhosphositePlus database and an additional 1,752 kinase-substrate pairs from the PhosphoNetwork database. We then applied the linear regression model as implemented in glm function in R to test for the relation between kinase and substrate phosphosite. The tests are independently conducted for cis and trans interactions in the cohort of 77 human breast cancer samples. For cis-interactions, we used kinase protein expression as the independent variable and each of the kinase's phosphosite level as the dependent variable. For trans-interactions, we used kinase phospho-protein expression and

the substrate's protein expression as independent variables and the substrate's phosphosite level as the dependent variables. For each kinase-substrate-phosphosite pair to be tested, we required both kinase protein/phosphoprotein expression and phosphosite phosphorylation to be observed in at least 10 samples in the respective datasets and the overlapped dataset. The resulting p values were adjusted using the Benjamini-Hochberg procedure to FDR.

We determined kinase-phosphosite pairs as validated if they showed P value under 0.05 and positive regression coefficients in the PDX cohort. For the kinase-phosphosite pairs showing top significant associations in the regression analysis, we calculated the average of phosphorylation level for each of the substrate phosphosite and protein expression of the kinase within each of 4 breast cancer subtypes (Basal, Her2, LumA and LumB) for display in Figure 4.2 and 3.3.

Kinase group and family enrichment analysis

To test for whether kinases showing significant *cis* or *trans* correlations are enriched in kinase groups and families, we applied one tailed Fisher's exact test under a null hypothesis that the odds ratio of associated kinases in the family are not greater. The universe of kinases for each of the *cis* and *trans* test was defined as the total tested kinase, and the 2-by-2 table is constructed by (1) whether the kinase belongs to the kinase group/family, and (2) whether the kinase has any significant correlations.

Structural and co-phosphorylation analysis

We used HotSpot3D[74] to generate pairwise linear and 3D distance between residues within 1,288 proteins with available PDB structure. The active sites are mapped based on data from the RSCB

PDB as of January 2017 (http://www.rcsb.org/pdb/home/home.do). Pearson's correlation coefficients and adjusted P values are then calculated for each pair of two phosphosites within these proteins. We limited this correlation analysis to pairs of phosphosites jointly observed in at least 5 samples in the breast cancer cohort. For linear examinations of association landscapes, the lolliplots were generated and modified using the PCGP protein painter (http://explore.pediatriccancergenomeproject.org/proteinPainter).

Druggable kinase-substrate pairs and cascades analysis

We compiled a list 76 druggable genes along with their respective drugs, from established public databases as previously described[120]. We then limit the analysis to the 68 genes with per-sample average RSEM value greater than 100. We searched for significantly associated kinase-substrate pairs where either the kinase or substrate belongs to the set of druggable genes. The score of each pair is calculated as the sum of standardized kinase and substrate phosphosite levels:

$$X_{kinase-substrate} = \frac{X_{kinase} - \mu_{kinase}}{\sigma_{kinase}} + \frac{X_{substrate} - \mu_{substrate}}{\sigma_{substrate}}$$

whereby μ is the mean and σ is the standard deviation.

To identify druggable outliers in the conventional method, we scanned for events at the somatic mutation, CNV, RNA and protein expression for each gene. CNV, RNA, and protein expression outliers are identified as the ones greater than 2 interquartile ranges (IQR) above median as previously described [83]. We then complemented this conventional single-event analysis by identifying outliers using the kinase-substrate score.

To define the druggable cascade of each of the druggable kinase, we extracted all significantly associated *cis* phosphosites and its first-degree correlated *trans* substrate phosphosites. We then expanded one level beyond and extracted additional *trans* phosphosites associated with the phosphoprotein level of the first-degree substrates. The resulting cascades were visualized using heatmap showing levels of each protein/phosphoprotein/phosphosite and a network diagram using Cytoscape [138].

Clinical and immune correlation analysis

We conducted association analysis between the score for each kinase-substrate pair and pathogical stage, survival, radiation therapy, and immune signatures adjusted for their PAM50 subtypes. For continuous variables, including pathogical stage and immune scores, we used a Gaussian linear regression. For whether the sample has gone through radition therapy, we used a logistic regression model. We used the Cox proportional Hazards model for survival analysis. The resulting p values were adjusted to FDR using the Benjamini-Hochberg procedure.

Conclusion

Revealing fundamental interactions between DNA, RNA and protein empowered research and application of molecular biology in the past century. Similarly, omics data at each level urgently require integration. I plan build a research program to further develop bioinformatics tools and integrate large-scale omics. Inference between genomics variants, transcriptome, proteome, PTMs and signaling networks will facilitate our understanding of each oncogenic event. Finally, their integration will help us understand the molecular dynamics of cancer at a comprehensive scale and inform personalized medicine.

Reference

1 Lichtenstein, P. *et al.* Environmental and heritable factors in the causation of cancer--analyses of cohorts of twins from Sweden, Denmark, and Finland. *The New England journal of medicine* **343**, 78-85, doi:10.1056/NEJM200007133430201 (2000).

2 Bodmer, W. & Tomlinson, I. in *Current Opinion in Genetics and Development* Vol. 20 262-267 (2010).

3 Rahman, N. Realizing the promise of cancer predisposition genes. *Nature* **505**, 302-308, doi:10.1038/nature12981 (2014).

4 Lu, C. *et al.* Patterns and functional implications of rare germline variants across 12 cancer types. *Nature communications* **6**, 10086, doi:10.1038/ncomms10086 (2015).

5 Southey, M. C. *et al.* PALB2, CHEK2 and ATM rare variants and cancer risk: data from COGS. *J Med Genet*, doi:10.1136/jmedgenet-2016-103839 (2016).

6 Zhang, J. *et al.* Germline mutations in predisposition genes in pediatric cancer. *New England Journal of Medicine* **373**, 2336-2346, doi:10.1056/NEJMoa1508054 (2015).

7 Amendola, L. M. *et al.* Performance of ACMG-AMP Variant-Interpretation Guidelines among Nine Laboratories in the Clinical Sequencing Exploratory Research Consortium. *Am J Hum Genet* **98**, 1067-1076, doi:10.1016/j.ajhg.2016.03.024 (2016).

8 Richards, S. *et al.* Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genetics in Medicine* **17**, 405-424, doi:10.1038/gim.2015.30 (2015).

9 Knudson, A. G. Mutation and cancer: statistical study of retinoblastoma. *Proceedings of the National Academy of Sciences of the United States of America* **68**, 820-823, doi:10.1073/pnas.68.4.820 (1971).

10 Knudson, A. G. Two genetic hits (more or less) to cancer. *Nat Rev Cancer* **1**, 157-162, doi:10.1038/35101031 (2001).

11 Green, E. D., Guyer, M. S. & National Human Genome Research, I. Charting a course for genomic medicine from base pairs to bedside. *Nature* **470**, 204-213, doi:10.1038/nature09764 (2011).

12 Garnett, M. J. *et al.* Systematic identification of genomic markers of drug sensitivity in cancer cells. *Nature* **483**, 570-575, doi:10.1038/nature11005 (2012).

13 Rubio-Perez, C. *et al.* In silico prescription of anticancer drugs to cohorts of 28 tumor types reveals targeting opportunities. *Cancer Cell* **27**, 382-396, doi:10.1016/j.ccell.2015.02.007 (2015).

14 Simon, R. & Roychowdhury, S. Implementing personalized cancer genomics in clinical trials. *Nat Rev Drug Discov* **12**, 358-369, doi:10.1038/nrd3979 (2013).

15 Chin, L., Andersen, J. N. & Futreal, P. A. Cancer genomics: from discovery science to personalized medicine. *Nat Med* **17**, 297-303, doi:10.1038/nm.2323 (2011).

16 Li, S. *et al.* Endocrine-therapy-resistant ESR1 variants revealed by genomic characterization of breast-cancer-derived xenografts. *Cell reports* **4**, 1116-1130, doi:10.1016/j.celrep.2013.08.022 (2013).

17 Tentler, J. J. *et al.* Patient-derived tumour xenografts as models for oncology drug development. *Nat Rev Clin Oncol* **9**, 338-350, doi:10.1038/nrclinonc.2012.61 (2012).

18 Ding, L. *et al.* Genome remodelling in a basal-like breast cancer metastasis and xenograft. *Nature* **464**, 999-1005, doi:10.1038/nature08989 (2010).

19 Holton, P. *et al.* Initial assessment of the pathogenic mechanisms of the recently identified Alzheimer risk Loci. *Annals of human genetics* **77**, 85-105, doi:10.1111/ahg.12000 (2013).

20 Brown, K. E. *et al.* Proteomic profiling of patient-derived glioblastoma xenografts identifies a subset with activated EGFR: implications for drug development. *J Neurochem* **133**, 730-738, doi:10.1111/jnc.13032 (2015).

21 Li, H. *et al.* Proteomic Characterization of Head and Neck Cancer Patient-Derived Xenografts. *Mol Cancer Res* **14**, 278-286, doi:10.1158/1541-7786.MCR-15-0354 (2016).

22 Comprehensive molecular portraits of human breast tumours. *Nature* **490**, 61-70, doi:10.1038/nature11412 (2012).

23 Mertins, P. *et al.* Proteogenomics connects somatic mutations to signalling in breast cancer. *Nature* **advance on**, doi:10.1038/nature18003 (2016).

24 Gujral, T. S. *et al.* Profiling phospho-signaling networks in breast cancer using reverse-phase protein arrays. *Oncogene* **32**, 3470-3476, doi:10.1038/onc.2012.378 (2013).

25 Akbani, R. *et al.* A pan-cancer proteomic perspective on The Cancer Genome Atlas. *Nature communications* **5**, 3887, doi:10.1038/ncomms4887 (2014).

26 Al-Ejeh, F. *et al.* Kinome profiling reveals breast cancer heterogeneity and identifies targeted therapeutic opportunities for triple negative breast cancer. *Oncotarget* **5**, 3145-3158, doi:10.18632/oncotarget.1865 (2014).

27 Zhang, H. *et al.* Integrated Proteogenomic Characterization of Human High-Grade Serous Ovarian Cancer. *Cell* **166**, 755-765, doi:10.1016/j.cell.2016.05.069 (2016).

28 Zhang, B. *et al.* Proteogenomic characterization of human colon and rectal cancer. *Nature* **513**, 382-387, doi:10.1038/nature13438 (2014).

29 Duncan, J. S. *et al.* Dynamic reprogramming of the kinome in response to targeted MEK inhibition in triple-negative breast cancer. *Cell* **149**, 307-321, doi:10.1016/j.cell.2012.02.053 (2012).

30 Fleuren, E. D., Zhang, L., Wu, J. & Daly, R. J. The kinome 'at large' in cancer. *Nat Rev Cancer* **16**, 83-98, doi:10.1038/nrc.2015.18 (2016).

31 Newman, R. H. *et al.* Construction of human activity-based phosphorylation networks. *Mol Syst Biol* **9**, 655, doi:10.1038/msb.2013.12 (2013).

32 Mashl, R. J. *et al.* GenomeVIP: a cloud platform for genomic variant discovery and interpretation. *Genome Res*, doi:10.1101/gr.211656.116 (2017).

33 McKenna, A. *et al.* The genome analysis toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research* **20**, 1297-1303, doi:10.1101/gr.107524.110 (2010).

34 Koboldt, D. C. *et al.* VarScan 2: Somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Research* **22**, 568-576, doi:10.1101/gr.129684.111 (2012).

35 Ye, K., Schulz, M. H., Long, Q., Apweiler, R. & Ning, Z. Pindel: A pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics* **25**, 2865-2871, doi:10.1093/bioinformatics/btp394 (2009).

36 Ye, K. *et al.* Systematic discovery of complex insertions and deletions in human cancers. *Nature medicine* **advance on**, 1-10, doi:10.1038/nm.4002 (2015).

37 Koire, A., Katsonis, P. & Lichtarge, O. Repurposing Germline Exomes of the Cancer Genome Atlas Demands a Cautious Approach and Sample-Specific Variant Filtering. *Pac Symp Biocomput* **21**, 207-218 (2016).

38 Landrum, M. J. *et al.* ClinVar: public archive of interpretations of clinically relevant variants. *Nucleic acids research* **44**, D862-868, doi:10.1093/nar/gkv1222 (2015).

39 McKusick-Nathans Institute of Genetic Medicine, J. H. U. B., MD). Online Mendelian Inheritance in Man, OMIM®. (2016).

40 Halavi, M., Maglott, D., Gorelenkov, V. & Rubinstein, W. MedGen. (2013).

41 INSERM. Orphanet: an online database of rare diseases and orphan drugs. (1997).

42 Lu, C. *et al.* Patterns and functional implications of rare germline variants across 12 cancer types. *Nature communications* **6**, 10086, doi:10.1038/ncomms10086 (2015).

43 Ballinger, M. L. *et al.* Monogenic and polygenic determinants of sarcoma risk: an international genetic study. *Lancet Oncol* **17**, 1261-1271, doi:10.1016/S1470-2045(16)30147-4 (2016).

44 Basu, S. & Pan, W. Comparison of statistical tests for disease association with rare variants. *Genet Epidemiol* **35**, 606-619, doi:10.1002/gepi.20609 (2011).

45 Schmidt, L. *et al.* Two North American families with hereditary papillary renal carcinoma and identical novel mutations in the MET proto-oncogene. *Cancer Research* **58**, 1719-1722 (1998).

46 Cui, W. *et al.* PIK3CA amplification and PTEN loss in diffused large B-cell lymphoma. *Oncotarget* **8**, 66237-66247, doi:10.18632/oncotarget.19889 (2017).

47 Lek, M. *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285-291, doi:10.1038/nature19057 (2016).

48 Chen, K. *et al.* Clinical actionability enhanced through deep targeted sequencing of solid tumors. *Clin Chem* **61**, 544-553, doi:10.1373/clinchem.2014.231100 (2015).

49 Krassowski, M. *et al.* ActiveDriverDB: human disease mutations and genome variation in post-translational modification sites of proteins. *Nucleic Acids Res*, doi:10.1093/nar/gkx973 (2017).

50 Hornbeck, P. V. *et al.* PhosphoSitePlus: a comprehensive resource for investigating the structure and function of experimentally determined post-translational modifications in man and mouse. *Nucleic acids research* **40**, D261-270, doi:10.1093/nar/gkr1122 (2012).

51 Patil, M., Pabla, N., Huang, S. & Dong, Z. Nek1 phosphorylates Von Hippel-Lindau tumor suppressor to promote its proteasomal degradation and ciliary destabilization. *Cell Cycle* **12**, 166-171, doi:10.4161/cc.23053 (2013).

52 Plaza-Menacho, I. *et al.* RET Functions as a Dual-Specificity Kinase that Requires Allosteric Inputs from Juxtamembrane Elements. *Cell Rep* **17**, 3319-3332, doi:10.1016/j.celrep.2016.11.061 (2016).

53 Kawamoto, Y. *et al.* Identification of RET autophosphorylation sites by mass spectrometry. *The Journal of biological chemistry* **279**, 14213-14224, doi:10.1074/jbc.M312600200 (2004).

54 Gabant, G. *et al.* Autophosphorylated residues involved in the regulation of human chk2 in vitro. *Journal of molecular biology* **380**, 489-503, doi:10.1016/j.jmb.2008.04.053 (2008).

55 Bose, R. *et al.* Activating HER2 mutations in HER2 gene amplification negative breast cancer. *Cancer Discovery* **3**, 224-237, doi:10.1158/2159-8290.CD-12-0349 (2013).

56 Jimenez, C. *et al.* A novel point mutation of the RET protooncogene involving the second intracellular tyrosine kinase domain in a family with medullary thyroid carcinoma. *J Clin Endocrinol Metab* **89**, 3521-3526, doi:10.1210/jc.2004-0073 (2004).

57 John, E. M. *et al.* Prevalence of pathogenic BRCA1 mutation carriers in 5 US racial/ethnic groups. *JAMA : the journal of the American Medical Association* **298**, 2869-2876, doi:10.1016/S0084-3954(08)79042-0 (2007).

58 Kurian, A. W. BRCA1 and BRCA2 mutations across race and ethnicity: distribution and clinical implications. *Curr Opin Obstet Gynecol* **22**, 72-78, doi:10.1097/GCO.0b013e328332dca3 (2010).

59 Cheng, D. T. *et al.* Comprehensive detection of germline variants by MSK-IMPACT, a clinical diagnostic platform for solid tumor molecular oncology and concurrent cancer predisposition testing. *BMC Med Genomics* **10**, 33, doi:10.1186/s12920-017-0271-4 (2017).

60 Parsons, D. W. *et al.* Diagnostic Yield of Clinical Tumor and Germline Whole-Exome Sequencing for Children With Solid Tumors. *JAMA Oncol*, doi:10.1001/jamaoncol.2015.5699 (2016).

61 Hilton, J. L. *et al.* Inactivation of BRCA1 and BRCA2 in ovarian cancer. *J Natl Cancer Inst* **94**, 1396-1406 (2002).

62 Morak, M. *et al.* Loss of MSH2 and MSH6 due to heterozygous germline defects in MSH3 and MSH6. *Fam Cancer*, doi:10.1007/s10689-017-9975-z (2017).

63 Nagy, E. & Maquat, L. E. in *Trends in Biochemical Sciences* Vol. 23 198-199 (1998).

64 Rivas, M. A. *et al.* Human genomics. Effect of predicted protein-truncating genetic variants on the human transcriptome. *Science* **348**, 666-669, doi:10.1126/science.1261877 (2015).

65 Reimand, J., Wagih, O. & Bader, G. D. Evolutionary constraint and disease associations of post-translational modification sites in human genomes. *PLoS Genet* **11**, e1004919, doi:10.1371/journal.pgen.1004919 (2015).

66 Reimand, J., Wagih, O. & Bader, G. D. The mutational landscape of phosphorylation signaling in cancer. *Sci Rep* **3**, 2651, doi:10.1038/srep02651 (2013).

67 Jansson, M. *et al.* Arginine methylation regulates the p53 response. *Nat Cell Biol* **10**, 1431-1439, doi:10.1038/ncb1802 (2008).

68 Tibbetts, R. S. *et al.* Functional interactions between BRCA1 and the checkpoint kinase ATR during genotoxic stress. *Genes & development* **14**, 2989-3002 (2000).

69 Robinson, J. T. *et al.* Integrative genomics viewer. *Nat Biotechnol* **29**, 24-26, doi:10.1038/nbt.1754 (2011).

70 Vogelstein, B. *et al.* Cancer genome landscapes. *Science (New York, N.Y.)* **339**, 1546-1558, doi:10.1126/science.1235122 (2013).

71 Zhou, W. *et al.* TransVar: a multilevel variant annotator for precision genomics. *Nat Methods* **12**, 1002-1003, doi:10.1038/nmeth.3622 (2015).

72 Kumar, P., Henikoff, S. & Ng, P. C. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat Protoc* **4**, 1073-1081, doi:10.1038/nprot.2009.86 (2009).

73 Adzhubei, I., Jordan, D. M. & Sunyaev, S. R. Predicting functional effect of human missense mutations using PolyPhen-2. *Curr Protoc Hum Genet* **Chapter 7**, Unit7 20, doi:10.1002/0471142905.hg0720s76 (2013).

74 Niu, B. *et al.* Protein-structure-guided discovery of functional mutations across 19 cancer types. *Nat Genet* **48**, 827-837, doi:10.1038/ng.3586 (2016).

75 Adzhubei, I. A. *et al.* A method and server for predicting damaging missense mutations. *Nat Methods* **7**, 248-249, doi:10.1038/nmeth0410-248 (2010).

76 McLaren, W. *et al.* The Ensembl Variant Effect Predictor. *Genome Biol* **17**, 122, doi:10.1186/s13059-016-0974-4 (2016).

77 Ruderfer, D. M. *et al.* Patterns of genic intolerance of rare copy number variation in 59,898 human exomes. *Nat Genet* **48**, 1107-1111, doi:10.1038/ng.3638 (2016).

78 Fromer, M. & Purcell, S. M. Using XHMM Software to Detect Copy Number Variation in Whole-Exome Sequencing Data. *Curr Protoc Hum Genet* **81**, 7 23 21-21, doi:10.1002/0471142905.hg0723s81 (2014).

79 Chatterjee, R. *et al.* Traditional and targeted exome sequencing reveals common, rare and novel functional deleterious variants in RET-signaling complex in a cohort of living US patients with urinary tract malformations. *Hum Genet*, doi:10.1007/s00439-012-1181-3 (2012).

80 Rauniyar, N. & Yates, J. R., 3rd. Isobaric labeling-based relative quantification in shotgun proteomics. *J Proteome Res* **13**, 5293-5309, doi:10.1021/pr500880b (2014).

81 Bondarenko, G. *et al.* Patient-Derived Tumor Xenografts Are Susceptible to Formation of Human Lymphocytic Tumors. *Neoplasia* **17**, 735-741, doi:10.1016/j.neo.2015.09.004 (2015).

82 Kanehisa, M. & Goto, S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res* **28**, 27-30 (2000).

83 Mertins, P. *et al.* Proteogenomics connects somatic mutations to signalling in breast cancer. *Nature* **534**, 55-62, doi:10.1038/nature18003 (2016).

84 Zhang, B. *et al.* Proteogenomic characterization of human colon and rectal cancer. *Nature* **513**, 382-387, doi:10.1038/nature13438 (2014).

85 Prat, A. & Perou, C. M. Deconstructing the molecular portraits of breast cancer. *Mol Oncol* **5**, 5-23, doi:10.1016/j.molonc.2010.11.003 (2011).

86 Sorlie, T. *et al.* Repeated observation of breast tumor subtypes in independent gene expression data sets. *Proc Natl Acad Sci U S A* **100**, 8418-8423, doi:10.1073/pnas.0932692100 (2003).

87 Croft, D. *et al.* The Reactome pathway knowledgebase. *Nucleic Acids Res* **42**, D472-477, doi:10.1093/nar/gkt1102 (2014).

88 Comprehensive molecular portraits of human breast tumours. *Nature* **490**, 61-70, doi:10.1038/nature11412 (2012).

89 Janakiraman, M. *et al.* Genomic and biological characterization of exon 4 KRAS mutations in human cancer. *Cancer Res* **70**, 5901-5911, doi:10.1158/0008-5472.CAN-10-0192 (2010).

90 Edkins, S. *et al.* Recurrent KRAS codon 146 mutations in human colorectal cancer. *Cancer Biol Ther* **5**, 928-932 (2006).

91 Paik, S., Kim, C. & Wolmark, N. HER2 status and benefit from adjuvant trastuzumab in breast cancer. *N Engl J Med* **358**, 1409-1411, doi:10.1056/NEJMc0801440 (2008).

92 Drebin, J. A., Link, V. C., Stern, D. F., Weinberg, R. A. & Greene, M. I. Down-modulation of an oncogene protein product and reversion of the transformed phenotype by monoclonal antibodies. *Cell* **41**, 697-706 (1985).

93 Carter, P. *et al.* Humanization of an anti-p185HER2 antibody for human cancer therapy. *Proc Natl Acad Sci U S A* **89**, 4285-4289 (1992).

94 Witters, L. *et al.* Synergistic inhibition with a dual epidermal growth factor receptor/HER-2/neu tyrosine kinase inhibitor and a disintegrin and metalloprotease inhibitor. *Cancer Res* **68**, 7083-7089, doi:10.1158/0008-5472.CAN-08-0739 (2008).

95 Medina, P. J. & Goodin, S. Lapatinib: a dual inhibitor of human epidermal growth factor receptor tyrosine kinases. *Clin Ther* **30**, 1426-1447, doi:10.1016/j.clinthera.2008.08.008 (2008).

96 Liu, P., Cheng, H., Roberts, T. M. & Zhao, J. J. Targeting the phosphoinositide 3-kinase pathway in cancer. *Nat Rev Drug Discov* **8**, 627-644, doi:10.1038/nrd2926 (2009).

97 Baselga, J. *et al.* Everolimus in postmenopausal hormone-receptor-positive advanced breast cancer. *N Engl J Med* **366**, 520-529, doi:10.1056/NEJMoa1109653 (2012).

98 Mayer, I. A. *et al.* Stand up to cancer phase Ib study of pan-phosphoinositide-3-kinase inhibitor buparlisib with letrozole in estrogen receptor-positive/human epidermal growth factor receptor 2-negative metastatic breast cancer. *J Clin Oncol* **32**, 1202-1209, doi:10.1200/JCO.2013.54.0518 (2014).

99 Bendell, J. C. *et al.* Phase I, dose-escalation study of BKM120, an oral pan-Class I PI3K inhibitor, in patients with advanced solid tumors. *J Clin Oncol* **30**, 282-290, doi:10.1200/JCO.2011.36.1360 (2012).

100 Zhang, H. *et al.* Patient-derived xenografts of triple-negative breast cancer reproduce molecular features of patient tumors and respond to mTOR inhibition. *Breast Cancer Res* **16**, R36, doi:10.1186/bcr3640 (2014).

101 Xu, S. *et al.* Combined targeting of mTOR and AKT is an effective strategy for basal-like breast cancer in patient-derived xenograft models. *Mol Cancer Ther* **12**, 1665-1675, doi:10.1158/1535-7163.MCT-13-0159 (2013).

102 Schnitt, S. J. Classification and prognosis of invasive breast cancer: from morphology to molecular taxonomy. *Mod Pathol* **23 Suppl 2**, S60-64, doi:10.1038/modpathol.2010.33 (2010).

103 Wang, M. L. *et al.* Targeting BTK with ibrutinib in relapsed or refractory mantle-cell lymphoma. *N Engl J Med* **369**, 507-516, doi:10.1056/NEJMoa1306220 (2013).

104 Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754-1760, doi:10.1093/bioinformatics/btp324 (2009).

105 Koboldt, D. C. *et al.* VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res* **22**, 568-576, doi:10.1101/gr.129684.111 (2012).

106 Koboldt, D. C., Larson, D. E. & Wilson, R. K. Using VarScan 2 for Germline Variant Calling and Somatic Mutation Detection. *Curr Protoc Bioinformatics* **44**, 15 14 11-15 14 17, doi:10.1002/0471250953.bi1504s44 (2013).

107 Koboldt, D. C. *et al.* VarScan: variant detection in massively parallel sequencing of individual and pooled samples. *Bioinformatics* **25**, 2283-2285, doi:10.1093/bioinformatics/btp373 (2009).

108 McCormick, R. F., Truong, S. K. & Mullet, J. E. RIG: Recalibration and interrelation of genomic sequence data with the GATK. *G3 (Bethesda)* **5**, 655-665, doi:10.1534/g3.115.017012 (2015).

109 Ye, K., Schulz, M. H., Long, Q., Apweiler, R. & Ning, Z. Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics* **25**, 2865-2871, doi:10.1093/bioinformatics/btp394 (2009).

110 Ding, L. *et al.* Genome remodelling in a basal-like breast cancer metastasis and xenograft. *Nature* **464**, 999-1005, doi:10.1038/nature08989 (2010).

111 Li, B. & Dewey, C. N. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* **12**, 323, doi:10.1186/1471-2105-12-323 (2011).

112 Wang, K. *et al.* MapSplice: accurate mapping of RNA-seq reads for splice junction discovery. *Nucleic Acids Res* **38**, e178, doi:10.1093/nar/gkq622 (2010).

113 Mertins, P. *et al.* Ischemia in tumors induces early and sustained phosphorylation changes in stress kinase pathways but does not affect global protein levels. *Mol Cell Proteomics* **13**, 1690-1704, doi:10.1074/mcp.M113.036392 (2014).

114 Ellis, M. J. *et al.* Connecting genomic alterations to cancer biology with proteomics: the NCI Clinical Proteomic Tumor Analysis Consortium. *Cancer Discov* **3**, 1108-1112, doi:10.1158/2159-8290.CD-13-0219 (2013).

115 Cox, J. *et al.* Accurate proteome-wide label-free quantification by delayed normalization and maximal peptide ratio extraction, termed MaxLFQ. *Mol Cell Proteomics* **13**, 2513-2526, doi:10.1074/mcp.M113.031591 (2014).

116 Tabb, D. L. *et al.* Reproducibility of Differential Proteomic Technologies in CPTAC Fractionated Xenografts. *J Proteome Res* **15**, 691-706, doi:10.1021/acs.jproteome.5b00859 (2016).

117 Mertins, P. *et al.* Integrated proteomic analysis of post-translational modifications by serial enrichment. *Nat Methods* **10**, 634-637, doi:10.1038/nmeth.2518 (2013).

118 Langfelder, P. & Horvath, S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* **9**, 559, doi:10.1186/1471-2105-9-559 (2008).

119 Cancer, T. & Atlas, G. Integrated genomic analyses of ovarian carcinoma. *Nature* **474**, 609-615, doi:10.1038/nature10166 (2011).

120 Huang, K. L. *et al.* Proteogenomic integration reveals therapeutic targets in breast cancer xenografts. *Nat Commun* **8**, 14864, doi:10.1038/ncomms14864 (2017).

121 Dinkel, H. *et al.* Phospho.ELM: a database of phosphorylation sites--update 2011. *Nucleic Acids Res* **39**, D261-267, doi:10.1093/nar/gkq1104 (2011).

122 Manning, G., Whyte, D. B., Martinez, R., Hunter, T. & Sudarsanam, S. The protein kinase complement of the human genome. *Science (New York, N.Y.)* **298**, 1912-1934, doi:10.1126/science.1075762 (2002).

123 Chartier, M., Chenard, T., Barker, J. & Najmanovich, R. Kinome Render: a stand-alone and web-accessible tool to annotate the human protein kinome tree. *PeerJ* **1**, e126, doi:10.7717/peerj.126 (2013).

124 Hornbeck, P. V. *et al.* PhosphoSitePlus: a comprehensive resource for investigating the structure and function of experimentally determined post-translational modifications in man and mouse. *Nucleic Acids Res* **40**, D261-270, doi:10.1093/nar/gkr1122 (2012).

125 Hu, J. *et al.* PhosphoNetworks: a database for human phosphorylation networks. *Bioinformatics* **30**, 141-142, doi:10.1093/bioinformatics/btt627 (2014).

126 Cross, D. A., Alessi, D. R., Cohen, P., Andjelkovich, M. & Hemmings, B. A. Inhibition of glycogen synthase kinase-3 by insulin mediated by protein kinase B. *Nature* **378**, 785-789, doi:10.1038/378785a0 (1995).

127 Chaikuad, A. *et al.* A unique inhibitor binding site in ERK1/2 is associated with slow binding kinetics. *Nat Chem Biol* **10**, 853-860, doi:10.1038/nchembio.1629 (2014).

128 Matsumoto, T. *et al.* Crystal structure of non-phosphorylated MAP2K6 in a putative auto-inhibition state. *J Biochem* **151**, 541-549, doi:10.1093/jb/mvs023 (2012).

129 Rossomando, A. J. *et al.* Identification of Tyr-185 as the site of tyrosine autophosphorylation of recombinant mitogen-activated protein kinase p42mapk. *Proc Natl Acad Sci U S A* **89**, 5779-5783 (1992).

130 Seger, R. & Krebs, E. G. The MAPK signaling cascade. *FASEB J* **9**, 726-735 (1995).

131 Yoshihara, K. *et al.* Inferring tumour purity and stromal and immune cell admixture from expression data. *Nat Commun* **4**, 2612, doi:10.1038/ncomms3612 (2013).

132 Sharma, K. *et al.* Ultradeep human phosphoproteome reveals a distinct regulatory nature of Tyr and Ser/Thr-based signaling. *Cell Rep* **8**, 1583-1594, doi:10.1016/j.celrep.2014.07.036 (2014).

133 Hochgrafe, F. *et al.* Tyrosine phosphorylation profiling reveals the signaling network characteristics of Basal breast cancer cells. *Cancer Res* **70**, 9391-9401, doi:10.1158/0008-5472.CAN-10-0911 (2010).

134 Croucher, D. R. *et al.* Involvement of Lyn and the atypical kinase SgK269/PEAK1 in a basal breast cancer signaling pathway. *Cancer Res* **73**, 1969-1980, doi:10.1158/0008-5472.CAN-12-1472 (2013).

135 Li, J. *et al.* TCPA: a resource for cancer functional proteomics data. *Nat Methods* **10**, 1046-1047, doi:10.1038/nmeth.2650 (2013).

136 The UniProt, C. UniProt: the universal protein knowledgebase. *Nucleic Acids Res* **45**, D158-D169, doi:10.1093/nar/gkw1099 (2017).

137 Hornbeck, P. V. *et al.* PhosphoSitePlus, 2014: mutations, PTMs and recalibrations. *Nucleic Acids Res* **43**, D512-520, doi:10.1093/nar/gku1267 (2015).

138 Shannon, P. *et al.* Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res* **13**, 2498-2504, doi:10.1101/gr.1239303 (2003).